Can attention-based transformers explain or interpret cyberbullying detection?

Kanishk Verma[1,2], Tijana Milosevic[1,2], Brian Davis[1]
ADAPT Centre[1], DCU Anti Bullying Centre[2],
Dublin City University, Ireland
kanishk.verma@adaptcentre.ie

Abstract

Automated textual cyberbullying detection is known to be a challenging task. It is sometimes expected that messages associated with bullying will either be a) abusive, b) targeted at a specific individual or group, or c) have a negative sentiment. Transfer learning by fine-tuning pre-trained attention-based transformer language models (LMs) has achieved near state-of-the-art (SOA) precision in identifying textual fragments as being bullying-related or not. This study looks closely at two SOA LMs, BERT and HateBERT, fine-tuned on real-life cyberbullying datasets from multiple social networking platforms. We intend to determine whether these finely calibrated pre-trained LMs learn textual cyberbullying attributes or syntactical features in the text. The results of our comprehensive experiments show that despite the fact that attention weights are drawn more strongly to syntactical features of the text at every layer, attention weights cannot completely account for the decision-making of such attention-based transformers.

1 Introduction

Repeated hostile and aggressive online behaviour to **intentionally** hurt or embarrass someone through digital communication technologies are generally understood as **Cyberbullying**. (Patchin and Hinduja, 2006). Recent research findings by (B et al., 2021) and (Milosevic, 2021), indicate that 44% children in 11 European countries and nearly 50% children in Ireland have reported an increase in cyberbullying during the COVID-19 lockdown restrictions across multiple social networking sites (SNS) and multiplayer online gaming (MOG) platforms. This growing amount of cyberbullying content emerging across multiple SNS and MOG platforms is alarming and necessitates more effective content moderation, as earlier studied by (Gillespie et al., 2020; Milosevic, 2018; Gillespie, 2018). Therefore, one crucial step toward efficient and effective content moderation is the ability to recognize and define the basis for automated cyberbullying detection systems to classify a textual expression or phrase as cyberbullying.

Recent computational cyberbullying research claim to have outstanding accuracy and precision in automating the identification of cyberbullying using state-of-the-art (SOA) deep learning algorithms like attention-based Transformers, Gated Recurrent Units (GRUs), Long-Short Term Memory (LSTMs). However, upon close examination of such research, including those by (Paul and Saha, 2020; Yadav et al., 2020; Behzadi et al., 2021; Tripathy et al., 2020; Pradhan et al., 2020; Fang et al., 2021) among others, reveal that they rely on datasets for *hate-speech* or *personal-attacks* by (Founta et al., 2018; Waseem and Hovy, 2016; Wulczyn et al., 2017) for cyberbullying identification. In reality, despite their inventive attempts, these studies can only determine whether a text is abusive or hateful. We consider it a poor decision to detect cyberbullying using such out-of-domain datasets.

Additionally, work by (Ruder et al., 2019; Howard and Ruder, 2018; Dodge et al., 2020) has demonstrated the efficacy of transfer learning by fine-tuning pre-trained deep layered language models (LMs) for a variety of natural language processing (NLP) tasks, such as text classification, thereby yielding impressive results. (Verma et al., 2022), have demonstrated that fine-tuning LMs like $BERT_{base-uncased}$ by (Devlin et al., 2018), and $Hate-BERT_{base-uncased}$ by (Caselli et al., 2020) outperform traditional machine learning algorithms and aid in more accurate detection of textual cyberbullying across multiple SNS platforms. Research by (Vaswani et al., 2017; Devlin et al., 2018) demonstrates that the attention-based mechanisms within the deeply layered architecture of such pre-trained LMs can display dependencies between input and output. High attention weights for

Proceedings of the 29th International Conference on Computational Linguistic, pages 16–29
October 12–17, 2022.

inputs (such as words) are frequently referred to be accountable for the output, which provides the model's interpretability (Mullenbach et al., 2018; Xie et al., 2017; Martins and Astudillo, 2016; Lei et al., 2017; Choi et al., 2016; Xu et al., 2015). To our knowledge, these assertions and presumptions have not undergone a formal evaluation for user-generated content (UGC) datasets collected from various SNS and MOG platforms categorically labelled for cyberbullying.

(Kitchin, 2017; Ananny and Crawford, 2018; Katzenbach and Ulbricht, 2019) question the existing opaqueness of automated algorithmic content moderation and decision-making practices by SNS and MOG platforms. It has thus become necessary to design and develop transparent and equitable algorithms for moderation and regulation. To that effect, we attempt to extend the work by (Verma et al., 2022) on multiple platform cyberbullying detection, by addressing the following research question,

- **RQ.1** Can attention-weights of attention-based LMs fine-tuned on real-life cyberbullying datasets be relied upon to detect and explain cyberbullying in an interpretable and understandable way?

Hence, we hypothesize that if attention-based LMs fine-tuned on real-life cyberbullying datasets learn textual cyberbullying traits for detecting cyberbullying; they would have higher attention weights for a) Parts-of-speech (POS) tags like adjectives, nouns, proper nouns, pronouns, and b) words with more negative sentiment. We also hypothesize that this assumption will be valid for text samples categorically annotated as cyberbullying across different datasets sourced from varied SNS and MOG platforms.

Content Warning: This article contains examples of abusive language in Section 5.4. All examples are taken from existing datasets (Section 3) to illustrate its composition.

2 Related Work

2.1 Cyberbullying Detection on Multiple platforms

There has been research on cross-platform cyberbullying detection, but they have had a narrow focus. (Edwards et al., 2020) devise a dataset from direct messages (SMS) shared between participants across multiple SNS platforms, social media posts collected from now-defunct Formspring.me [1] and tweets from Twitter[2] focusing only on one topic (2016 USA elections). However, despite their novel attempts at developing a cross-platform cyberbullying dataset and devising supervised machine learning classifiers to identify cyberbullying, their focus on a specific type of text-based communication like SMS and only on two types of SNS platforms, of which one is now defunct. On the other hand, (Nikhila et al., 2020; Yi and Zubiaga, 2022) also devise novel techniques to identify textual cyberbullying using adversarial neural network algorithms. Nevertheless, for training the classifiers, they rely on datasets by (Waseem and Hovy, 2016; Wulczyn et al., 2017) marked for either personal attacks or hate speech. On the contrary, work by (Van Bruwaene et al., 2020) is both novel and apt for cyberbullying research. They devise a high-quality dataset and experiment with Support Vector Machines (SVM), Convolutional neural networks (CNNs), and XGBOOST algorithms to develop a cross-platform cyberbullying detection system. To our knowledge, the work by (Van Bruwaene et al., 2020) is the only one that leverages real-life cyberbullying datasets. However, due to proprietary reasons, it is not yet made publicly available. (Verma et al., 2022) leverage real-life cyberbullying datasets collected by computational researchers from multiple SNS and MOG platforms such as Instagram[3], Twitter, ASK.fm[4], now-defunct SNS platforms Formspring.me, and Vine[5]. On training multiple binary cyberbullying classifiers on single platforms and benchmarking their efficacy on different platforms, they found that attention-based LMs could achieve better precision and recall than traditional machine learning algorithms at classifying cyberbullying samples as cyberbullying. However (Verma et al., 2022) were unable to determine why these phenomena occur, and were also unable to establish whether the attention-based LMs were dependent on any textual cyberbullying traits (eg. profanities or negative sentiment words).

[1] an anonymous question-answering SNS

[2] https://twitter.com

[3] https://www.instagram.com

[4] ASK.fm - https://ask.fm; is an anonymous question-answering SNS platform

[5] Video-sharing platform like TikTok https://en.wikipedia.org/wiki/Vine_(service)

2.2 Analysing Attention in attention-based language models

Attention-based transformer LMs developed by (Devlin et al., 2018; Yang et al., 2019; Caselli et al., 2020) consists of a deep architecture with many hidden layers stacked on top of one another. Within these layers are many attention-heads or sub-layers that assign attention-weights to a token (word) for learning the importance of the token. Substantial research conducted by (Zhang et al., 2019; Adadi and Berrada, 2018; Sundararajan et al., 2017) and others have demonstrated frameworks for explaining and interpreting these deep-layered LMs by analyzing these attention-weights at every layer. Moreover, (Vig, 2019a; Vig and Belinkov, 2019) have developed tools and resources that aid in visualizing the attention weights. This allows human users to comprehend and trust the results of such deep-layered LMs. However, studies by (Jain and Wallace, 2019; Serrano and Smith, 2019; Sun and Lu, 2020; Vashishth et al., 2019) have demonstrated that these attention-based mechanisms solely cannot be relied upon for interpreting and explaining the intricate workings of LMs. The work by (Elsafoury et al., 2021) for interpreting the attention mechanism of BERT for cyberbullying is closest to our research. We thank the authors (Elsafoury et al., 2021) for their contributions and for making the code repository reproducible. However, they a) rely on out-of-domain datasets like hate speech and personal attack datasets and b) lack in-depth analysis of LMs decision-making for both textual cyberbullying and non-cyber bullying samples. Moreover, they do not report the attention-based LM(s) interpretation for real-world binary instances of cyberbullying in text.

3 Datasets

To overcome current dataset-related gaps in cyberbullying research, we select datasets that are a) annotated by either cyberbullying domain experts or b) clear and precise annotation guidelines aided the annotation for cyberbullying. To our knowledge, there are only *seven* real-life datasets in English language that have been devised for cyberbullying markers with such annotations. We categorise the seven datasets into four groups based on a) type of SNS or MOG platform and b) average length of tokens observed from each of the seven platforms (See Figure 1). These groups include,

- **Question-answering SNS**: Question-answering SNS are both anonymous and non-anonymous platforms like ASK.fm, Reddit[6], and Quora[7] that allow platform users to respond to questions posted by other users. Dataset devised by (Van Hee et al., 2018) from the ASK.fm platform is available in both English and Dutch. Annotations in this dataset are both binary and fine-grained, i.e., it is annotated for different cyberbullying forms and varied cyberbullying participant roles. (Reynolds et al., 2011) collected English language dataset from now-defunct Formspring.me. Annotations in this dataset are binary, i.e., textual samples are labelled as cyberbullying and non-cyber bullying.

- **Twitter SNS**: (Xu et al., 2012) formulated a dataset by collecting tweets[8] from Twitter in the English language. Their dataset annotations are annotated as binary textual cyberbullying samples and for varied author roles such as victim, bully, reporter, and others. (Salawu et al., 2020) formulated a dataset from tweets in the English language. They have various labels such as profanity, insult, spam, sarcasm, threat, exclusion, and bullying.

- **User-comment SNS**: User-comment SNS are platforms that allow users to comment on images or videos posted by other platform users. Such platforms include but are not limited to Instagram, Facebook, TikTok, Vine, etc. (Hosseinmardi et al., 2015) collected multi-modal data (inclusive of images and textual comments) from Instagram. The annotations in this dataset are for both cyberbullying and cyber aggression. (Rafiq et al., 2015) also collected multi-modal data (inclusive of videos and textual comments) from now-defunct Vine platform. The annotation in this dataset includes both cyberbullying and cyber-aggression.

- **MOG platforms**: On MOG platforms, players communicate on forums, in-game chats, or via voice (either in-built or by plug-in voice-call platforms like Discord). (Bretschneider and Peters, 2016) collected text from fo-

[6] https://www.reddit.com
[7] https://www.quora.com
[8] https://help.twitter.com/en/resources/twitter-guide/topics/how-to-join-the-conversation-on-twitter/how-to-tweet

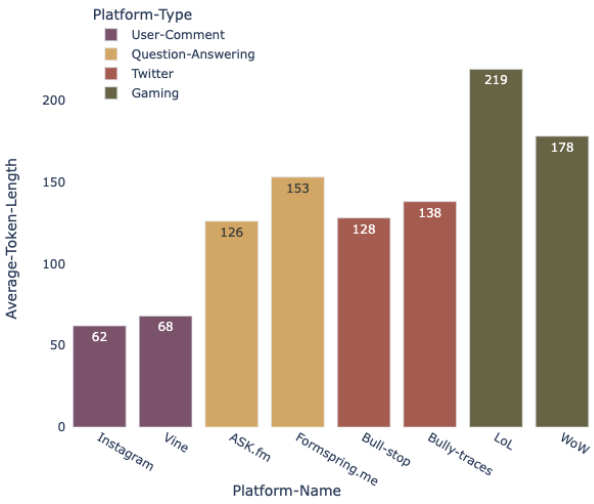

Figure 1: Length of tokens in every dataset

rums of highly popular MOG platforms like like World-of-Warcraft (WoW)[9] and League-of-Legends (LoL)[10]. The annotations in this dataset are of two types a) role-based binary annotations, i.e., bully and victim, and b) binary labels for cyberbullying for each text sample.

As seen in Table 1, number of sentences for cyberbullying content is very low across all seven datasets. Each of the four grouped datasets were split into training, validation, and test sets. The training set makes up 80% for each of the grouped dataset, while the validation set and test set each make up for 10% of the datasets. The test set was further broken into two parts a) 90% was used to evaluate performance of fine-tuned LMs performed, and b) 10% was used to analyze attention weights and gradient-based feature importance scores for each layer and head in the architecture of the LMs.

4 Experiment Setup

Tasks depicted in the first block of Figure 7 (See Appendix A.3) are described in detail in sections 4.1 and 4.3. As depicted in other two blocks of the Figure 7, we discuss the strategies for fine-tuning, hyper-parameter optimization strategies, and the attention-weight and gradient-based feature importance score at each layer for both bullying and non-bullying sentences in Sections 4.2 and 4.4. The code repository for reproducing this study can be found online[11].

4.1 Data anonymization, pre-processing and handling data imbalance

4.1.1 Data Anonymization

Adhering to General Data Protection Regulation (GDPR) directive (Council of European Union, 2016), we fully anonymised and normalised the datasets for any Personally Identifiable Information (PII) data. Such data included but was not limited to email-address, user names, geographical locations, and user-profile details, among others. Using GATE Cloud (Tablan et al., 2013) and the TwitIE API (K. Bontcheva, 2013), PII data was de-identified by masking and replacing the original words with masked value. For example, the sentence *"mary@gmail.com is based in London"*, was masked as *"email-address is based in location"*.

4.1.2 Pre-processing

Due to the abundance of non-standard language in the datasets, including lexical variants like *supa* $\rightarrow$ *super*, and acronyms, e.g., *tbh* $\rightarrow$ *to be honest*, and spelling errors, we applied several normalization heuristics for spelling and slang corrections. We removed a) URLs, user mentions, and non-ASCII characters for all datasets, b) retweet (RT) markers in text for *twitter* datasets, and c) lower-cased all text, and d) converted contractions to formal format. We also gathered a list of slang words and acronyms with their standardized forms from an online website[12]. Finally, we developed an algorithm (See Appendix A.1 for details) to fix spelling errors with the most accurate semantic corrections.

4.1.3 Data Imbalance

The percentage of cyberbullying content in Table 1 shows a high imbalance skewed towards the non-bullying class. Handling the imbalance was paramount to avoid learning the *bias* towards the majority class in imbalanced datasets. Due to the limited nature of the dataset and to avoid the risk of losing context and sequence of words in a sentence, we leveraged the simple random oversampling technique (Moreo et al., 2016) over Synthetic Minority Oversampling Technique (SMOTE) (Bunkhumpornpat et al., 2009). It is worth noting

[9] https://us.forums.blizzard.com/en/wow/

[10] https://www.leagueoflegends.com/en-us/news/community/

[11] https://gitlab.computing.dcu.ie/vermak3/xai-cyberbullying-attention

[12] https://www.webopedia.com/reference/text-abbreviations/

Platform-Type	Study	% Cyberbullying Content	# of Sentences
Question-Answering	(Van Hee et al., 2018)	4.73	113,698
	(Reynolds et al., 2011)	9.42	25,802
User-Comment	(Hosseinmardi et al., 2015)	41.28	32,074
	(Rafiq et al., 2015)	34.58	78,249
Twitter	(Xu et al., 2012)*	5.99	9,965
	(Salawu et al., 2020)*	4.67	4,009
Gaming	(Bretschneider and Peters, 2016)	2.3	34,229

Table 1: Dataset Description.
"*" Numbers vary to original dataset, as the tweet is unavailable due to deletion by tweet authors.

that the random oversampling was done in only one training set, and data imbalance was not handled in the validation and test set to match real-life scenarios. Also, to verify if over-sampling techniques affect the classification models' accuracy, we run experiments with imbalanced and over-sampled datasets.

4.2 Language Models and Hyper-parameters

To ascertain which SOA LMs is able to a) better capture dependencies and b) learn better representation of cyberbullying text from noisy UGC data, we leverage pre-trained BERT$_{base-uncased}$ by (Devlin et al., 2018), and Hate-BERT$_{base-uncased}$ by (Caselli et al., 2020). BERT$_{base-uncased}$ is a bi-directional auto-encoding attention-based transformer with twelve layered transformer blocks, with each block containing twelve self-attention layers and a total of 768 hidden layers, resulting in approximately 110 M parameters. Hate-BERT$_{base-uncased}$ is a BERT LM re-trained on hateful comments from RAL-E Reddit's banned communities (Chandrasekharan et al., 2017). We utilized the implementation provided by HuggingFace's Transformer Library (Wolf et al., 2019) and by (Caselli et al., 2020), and follow (Verma et al., 2022) experiments to fine-tune the pre-trained LMs. To find optimal hyper-parameters, we used the Weights & Biases (Biewald, 2020) plug-ins to conduct multiple grid-based experiments with a varied range of hyper-parameters and optimized it to achieve maximum validation accuracy. The range of hyper-parameters includes,

- Maximum Token Length(s): `[128, 256]`
- Batch-size(s): `[8, 16, 32]`
- Epochs: `[2, 3, 4]`
- Loss Function: *Binary Cross Entropy*
- Optimizer Function: *Adam Weighted*
- Learning Rate(s): $1e^{-5}$, $2e^{-5}$, $3e^{-5}$, $4e^{-5}$, $5e^{-5}$

4.3 Collecting Parts-of-speech (POS) Tags & Sentiment Scores

To formally evaluate our hypothesis and assumptions addressed in Section 1. As the datasets leveraged in this study are a) sourced from SNS and MOG platforms, b) are not in formal language, and c) do not include POS tags or sentiment scores, we leveraged Spacy's POS tagger [13] (Honnibal and Montani, 2017) to collect POS tags and VADER by (Hutto and Gilbert, 2014) to collect sentiment scores. Please note that both POS tags and sentiment scores were collected only for 10% of the test-set, (See Table 4 in Appendix A.2).

4.4 Attention-weights and gradient-based importance scores

To address our **RQ.1** and compare with other experiments (Jain and Wallace, 2019; Serrano and Smith, 2019; Sun and Lu, 2020; Vashishth et al., 2019), we extract attention-weights of the fine-tuned LM(s) on 10 % of test-set reserved for attention analysis. Many experiments on transformer-based attention-analysis refer to gradient-based feature importance scores as a measure for providing importance of individual features with known semantics (Clark et al., 2019; Serrano and Smith, 2019; Sun and Lu, 2020). We leveraged the *Integrated Gradients* algorithm by (Sundararajan et al., 2017) for pytorch[14] to model interpretability by (Kokhlikyan et al., 2020) to compute the gradient-based feature importance scores on 10% of the test-set reserved for attention analysis. As the pre-trained LMs used

[13] https://spacy.io/usage/linguistic-features#pos-tagging

[14] A python language framework for deep learning https://pytorch.org/

in this study have 12 transformer block layers and 12 attention heads, we computed the mean attention weights for each head of every layer. Extending (Jain and Wallace, 2019)'s work, we use Pearson's correlation coefficient (PCC) to measure the linear correlation between mean importance scores and mean attention weights. Moreover, we also a) compute mean-attention weights and gradient-based feature importance scores for every token for every POS tag of the reserved test-set and b) observe both mean-attention weights and gradient-based feature importance scores for tokens in the reserved test-set that have a greater negative sentiment.

5 Results

5.1 Impact of Data Imbalance

To assess if our simple oversampling strategy on training data referred to in Section 4.1.3 yields any significant improvement over the no-sampling strategy on training data, we check the model's validation accuracy on the non-sampled validation dataset. As depicted in Table 5 (See Appendix A.4) , we find no significant differences in the performance of either oversampling or no-sampling when both BERT and HateBERT LMs are fine-tuned on *user-comment*, *twitter*, and *question-answering* datasets, except for the *gaming* dataset. Both LMs fine-tuned on over-sampled *gaming* dataset perform better than fine-tuning on no-sampled datasets. We believe this is because, as seen in Table 1, there are only 2.3% bullying samples in the *gaming* dataset, and it is highly skewed towards non-bullying samples.

5.2 Hyper-parameters Finetuning

As discussed in Section 4.2, we experiment with different combinations of hyper-parameters with the help of the Weights & Biases plug-in for grid-based experiments. Table 2 represents the results of optimal hyper-parameters on validation-set for both fine-tuned LMs on each of the four datasets. We find that hyper-parameters vary for each model on every dataset. Overall, optimal hyper-parameters include, token-length of 128, batch-sizes ranging from 8 to 32, learning-rates between 1e-5, 2e-5, 3e-5, and 5e-5, and with 2 Epochs. With the help of these hyper-parameters, maximum accuracy can be achieved on validation sets.

5.3 Cyberbullying Detection

After training and validating both $BERT_{base-uncased}$ and $HateBERT_{base-uncased}$ with the optimal hyper-parameters (See Table 2) for every dataset, we assessed both LMs performance for their F1-scores for a) bullying, and b) non-bullying samples. In cyberbullying detection, false negatives and false positives are crucial, especially in cases of imbalanced data. We believe that F1 scores for each class are an apt metric for evaluating classifiers. As depicted in the Table 3 fine-tuning the HateBERT LM for each of the four platform datasets, does perform better than just fine-tuning the BERT LM. Moreover, these generalized LMs perform better with the grouped Twitter datasets.

5.4 Attention-weights & Gradient-based feature analysis

5.4.1 Correlation between attention-weights & gradient-based feature importance scores

As observed in the Figure 2, the Pearson's correlation coefficient (PCC) between attention-weights and gradient-based feature importance scores for fine-tuned HateBERT ranges from *0.0129* for bullying-samples in *user-comment* dataset to *0.1202* for bullying-samples in *gaming* datasets. Whereas, for fine-tuned BERT the PCC between attention-weights and gradient-based feature importance scores ranges from 0.0042 for bullying-samples in *question-answering* datasets to 0.19 in *twitter dataset*. Overall, as depicted in the Figure 2, this PCC is close to zero for both BERT and HateBERT LMs fine-tuned on *user-comment* dataset. For BERT LM fine-tuned on *twitter* and *gaming* datasets, the PCC between attention-weights and gradient-based feature importance scores is in the range of 0.08 to 0.19. However, for HateBERT LM fine-tuned on *twitter* datasets, this is not the case; the PCC between attention-weights and gradient-based feature importance for this data is nearly zero (0.070-0.078).

From Table 3, we can deduce that HateBERT LM fine-tuned on *twitter* datasets has better F-scores than BERT LM fine-tuning on the same dataset. The near zero-correlation observed between mean attention-weights and gradient-based importance scores for both generalized LMs, especially for better performing HateBERT LM fine-tuned on *twitter* datasets, helps us substantiate

Model	Dataset	Token length	Batch-size	Learning Rate	Epochs	Val-Accuracy
BERT+	Gaming	128	8	$5e^{-5}$	4	0.7746
	UC	128	32	$3e^{-5}$	2	0.7476
	QA	128	8	$5e^{-5}$	2	0.9496
	Twitter	128	32	$2e^{-5}$	2	0.94
HateBERT+	Gaming	128	8	$1e^{-5}$	2	0.7478
	UC	128	16	$3e^{-5}$	2	0.7497
	QA	128	16	$1e^{-5}$	2	0.9498
	Twitter	128	32	$5e^{-5}$	2	0.939

Table 2: Optimal Hyper-parameters for every model with every dataset
Note: UC → user-comment; QA → question-answering

Model	Dataset	Bullying F1	Non-Bullying F1	Average F1
BERT+	QA	0.62	0.68	0.65
	UC	0.65	0.77	0.71
	Twitter	0.75	0.79	0.77
	Gaming	0.68	0.79	0.72
HateBERT+	QA	0.73	0.73	0.73
	UC	0.68	0.84	0.76
	Twitter	0.78	0.84	0.81
	Gaming	0.74	0.78	0.76

Table 3: Cyberbullying Classification Results
Note: UC → user-comment; QA → question-answering

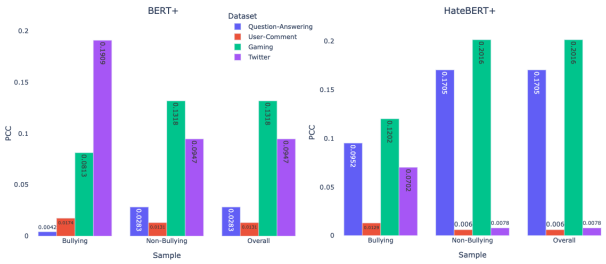

Figure 2: Pearson's Correlation between Mean Attention Weights and Mean Gradient-Importance Score for all LMs and Datasets

the claims by (Jain and Wallace, 2019; Serrano and Smith, 2019; Sun and Lu, 2020; Vashishth et al., 2019). They claim that while attention-mechanisms improve classification performance, *relying on attention-weights for interpretation is questionable at best*, holds true even for real-life SNS and MOG cyberbullying datasets.

5.4.2 Layer-wise attention Analysis

In this section, we examine the mean-attention weights at each layer for each POS tag as well as sentences with both stronger negative & positive sentiment that were taken from fine-tuned LMs that had a) a higher positive correlation between the mean attention weights and the mean gradient-based feature importance scores and b) a higher classification F-1 score. These models are BERT+$_{t}witter$ and HateBERT+$_{gaming}$ as depicted in the Table 3 and the Figure 2. So, using data from both Twitter and gaming datasets, we provide layer-wise analysis as follows,

- **Layer-wise Attention for Parts-of-speech Tags &**
  In Figures 3 and 4, we represent mean-attention attention weights at each layer for every POS tag in the *twitter* and *gaming* datasets. For adjectives in both bullying and non-bullying samples in these datasets, fine-tuned BERT and HateBERT models have mean attention weights ranging from 0.051 to 0.062. For verbs in bullying samples, fine-tuned BERT has 0.1 mean attention weight at layer 6, and at the end of layer 12, it drops to 0.09, whereas in the fine-tuned HateBERT model for bullying samples, the mean attention weight is as low as it is for adjectives. For nouns in bullying samples, fine-tuned BERT has a mean attention weight of 0.051, and fine-tuned HateBERT has a mean attention weight of 0.14 at the starting layers, but by layers 11 and 12, it drops down to 0.074. For proper nouns, fine-tuned BERT and HateBERT have a much higher mean attention weight for bullying samples. However, in non-bullying samples, fine-tuned HateBERT has a lower mean attention weight of 0.051 throughout all layers. This, in a way, disproves our hypothesis that words that are adjectives, verbs, nouns, and proper nouns will have higher mean attention weights. As depicted in Figures 3 and 4, mean

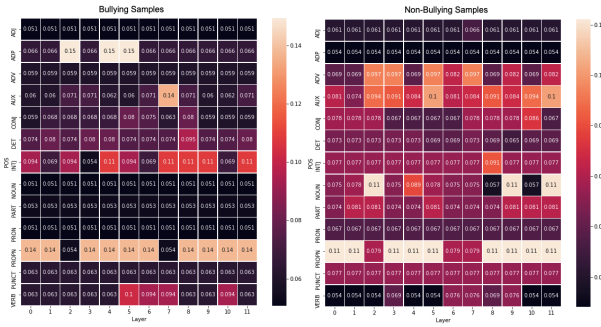

Figure 3: BERT+ Mean-attention weights per Layer POS-tag wise

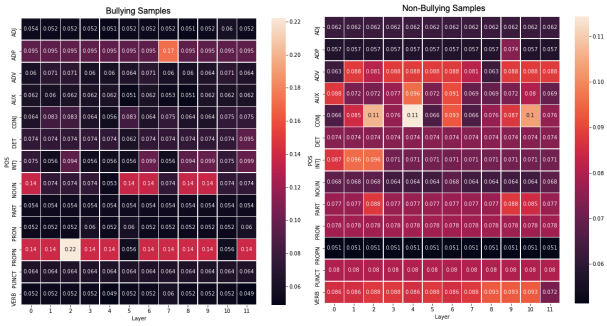

Figure 4: HateBERT+ Mean-attention weights per Layer POS-tag wise

attention weights for auxiliaries, determiners, ad-positions, conjunctions, and interjections in comparison to adjectives, verbs and nouns are consistently higher at almost all layers for fine-tuned BERT and HateBERT models in both bullying and non-bullying samples.

- **Negative Words & Attention**
  As discussed in an earlier section , we assume that negative sentiment words will have greater mean attention weights, so we compare these weights with the mean attention weights of positive sentiment words. Also, as mentioned in Section 4.3, we leveraged Vader to generate sentiment scores for all tokens in the reserved test set. We selected four words with greater negative sentiment (f*ck, a*s, b*tch, stu*id) and four words with greater positive sentiment (like, cute, pretty, truth). The Figures 5 and 6 represent mean-attention weights per layer for each of the words. We find that the positive sentiment word *pretty* has a similar mean attention weight as the negative sentiment words *b*tch and a*s* at all layers for both fine-tuned BERT and HateBERT models. This disproves our second assumption that negative sentiment words will have greater mean attention weights. For brevity, our analysis on two sample sentences at every layer and head are briefly discussed in Appendix A.5.

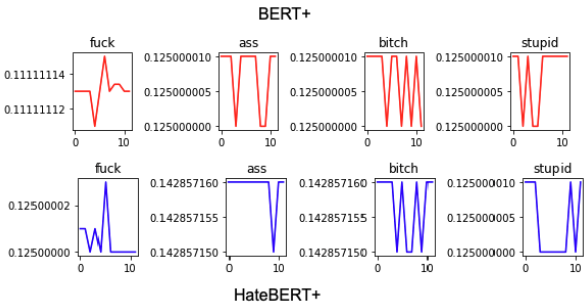

Figure 5: Attention per layer for Negative Sentiment words

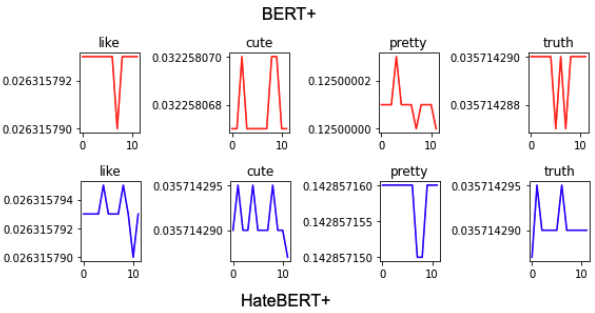

Figure 6: Attention per layer for Positive Sentiment words

6 Conclusion & Future Work

Our comprehensive experiments using two LMs and four grouped real-life cyberbullying datasets from various SNS and MOG platforms revealed that: a) a fine-tuned HateBERT is better at classifying cyberbullying samples as cyberbullying; b) there is almost no correlation between the attention-weights and gradient-based feature importance scores; and c) an attention-based transformer model tuned for cyberbullying classification relies more on syntactical features in text, and d) attention-based transformer model fine-tuned for cyberbullying classification have similar "attention" for both negative and positive sentiment words. While our experiments show that to some level, attention-based transformers fine-tuned on real-life cyberbullying datasets do aid in interpreting and explaining its decision-making based on cyberbullying features, there is still a lot more work to be done in devising transparent and fair LM(s) for cyberbullying

detection. While we demonstrate comprehensive methods to interpret and explain fine-tuned LMs on real-life SNS and MOG text cyberbullying classification, we acknowledge that due to the diverse forms and roles of cyberbullying, our work is limited by binary cyberbullying categories. Due to the current paucity of fine-grained cyberbullying datasets, in the future, we will attempt to use the learned representation of these fine-tuned LM(s) on fine-grained pre-adolescent datasets by (Sprugnoli et al., 2018).

7 Acknowledgements

We thank the authors (Van Hee et al., 2018; Reynolds et al., 2011; Xu et al., 2012; Salawu et al., 2020; Hosseinmardi et al., 2015; Rafiq et al., 2015; Bretschneider and Peters, 2016) for sharing the dataset. The research conducted in this publication was funded by the Irish Research Council and Google, Ireland, under grant number EPSPG/2021/161.

References

Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, 20(3):973–989.

Lobe B, Velicu A, Staksrud E, Chaudron S, and Di Gioia R. 2021. How children (10-18) experienced online risks during the covid-19 lockdown - spring 2020. (KJ-NA-30584-EN-N (online),KJ-NA-30584-EN-C (print)).

Mitra Behzadi, Ian G Harris, and Ali Derakhshan. 2021. Rapid cyber-bullying detection method using compact bert models. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 199–202. IEEE.

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Uwe Bretschneider and Ralf Peters. 2016. Detecting cyberbullying in online communities.

Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. 2009. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining*, pages 475–482, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2020. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. *CoRR*, abs/1906.04341.

Council of European Union. 2016. Regulation (eu) 2016/679 of the european parliament and of the council (general data protection regulation). https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *CoRR*, abs/2002.06305.

April Edwards, David Demoll, and Lynne Edwards. 2020. Detecting cyberbullying activity across platforms. In *17th International Conference on Information Technology–New Generations (ITNG 2020)*, pages 45–50. Springer.

Fatma Elsafoury, Stamos Katsigiannis, Steven R Wilson, and Naeem Ramzan. 2021. Does bert pay attention to cyberbullying? In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1900–1904.

Yong Fang, Shaoshuai Yang, Bin Zhao, and Cheng Huang. 2021. Cyberbullying detection in social networks using bi-gru with self-attention mechanism. *Information*, 12(4):171.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Tarleton Gillespie. 2018. *Custodians of the Internet*. Yale University Press.

Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T Roberts, Aram Sinnreich, and Sarah Myers West. 2020. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review*, 9(4):Article–number.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *CoRR*, abs/1402.3722.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer.

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

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

A. Funk M.A. Greenwood D. Maynard N. Aswani K. Bontcheva, L. Derczynski. 2013. witie: An open-source information extraction pipeline for microblog text. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*.

Christian Katzenbach and Lena Ulbricht. 2019. Algorithmic governance. *Internet Policy Review*, 8(4):1–18.

Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information, communication & society*, 20(1):14–29.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Tao Lei et al. 2017. *Interpretable neural models for natural language processing*. Ph.D. thesis, Massachusetts Institute of Technology.

André F. T. Martins and Ramón Fernandez Astudillo. 2016. From softmax to sparsemax: A sparse model of attention and multi-label classification. *CoRR*, abs/1602.02068.

Laffan D. O'Higgins Norman J Milosevic, T. 2021. Kidicoti: Kids' digital lives in covid-19 times: A study on digital practices, safety and wellbeing; key findings from ireland. https://antibullyingcentre.b-cdn.net/wp-content/uploads/2021/12/Short-report_Covid_for-media_TM_with-Author-names-1-2.pdf.

Tijana Milosevic. 2018. *Protecting children online?: Cyberbullying policies of social media companies*. The MIT Press.

Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. 2016. Distributional random oversampling for imbalanced text classification. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 805–808, New York, NY, USA. Association for Computing Machinery.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Munipalle Sai Nikhila, Aman Bhalla, and Pradeep Singh. 2020. Text imbalance handling and classification for cross-platform cyber-crime detection using deep learning. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Justin W Patchin and Sameer Hinduja. 2006. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth violence and juvenile justice*, 4(2):148–169.

Sayanta Paul and Sriparna Saha. 2020. Cyberbert: Bert for cyberbullying identification. *Multimedia Systems*, pages 1–8.

Ankit Pradhan, Venu Madhav Yatam, and Padmalochan Bera. 2020. Self-attention for cyberbullying detection. In *2020 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA)*, pages 1–6. IEEE.

Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 617–622. IEEE.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Semiu Salawu, Yulan He, and Jo Lumsden. 2020. Bullstop: A mobile app for cyberbullying prevention. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 70–74.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428, Online. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Valentin Tablan, Ian Roberts, Hamish Cunningham, and Kalina Bontcheva. 2013. Gatecloud.net: a platform for large-scale, open-source text processing on the cloud. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1983):20120071.

Jatin Karthik Tripathy, S Sibi Chakkaravarthy, Suresh Chandra Satapathy, Madhulika Sahoo, and V Vaidehi. 2020. Albert-based fine-tuning model for cyberbullying analysis. *Multimedia Systems*, pages 1–9.

David Van Bruwaene, Qianjia Huang, and Diana Inkpen. 2020. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54(4):851–874.

Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2018. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Kanishk Verma, Tijana Milosevic, Keith Cortis, and Brian Davis. 2022. Benchmarking language models for cyberbullying identification and classification from social-media texts. In *Proceedings of The First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 26–31, Marseille, France. European Language Resources Association.

Jesse Vig. 2019a. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Jesse Vig. 2019b. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Qizhe Xie, Xuezhe Ma, Zihang Dai, and Eduard Hovy. 2017. An interpretable knowledge transfer model for knowledge base completion. *arXiv preprint arXiv:1704.05908*.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Jaideep Yadav, Devesh Kumar, and Dheeraj Chauhan. 2020. Cyberbullying detection using pre-trained bert model. In *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pages 1096–1100. IEEE.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Peiling Yi and Arkaitz Zubiaga. 2022. Cyberbullying detection across social media platforms via platform-aware adversarial encoding.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*.

A Appendix

A.1 Data Normalisation Algorithm

As SNS and MOG text is short, an incorrect replacement of misspelt words can make the sentence lose its context. For example, if in the sentence *"i got a tkn to play" "tkn"* is replaced as *"taken"* instead of *"token"*, the sentence will lose its meaning. So, to avoid such an incorrect spell correction, it is important to understand the context of the sentence. To that effect we developed an algorithm 1 to fix spelling spelling errors with the most accurate semantic corrections by leveraging the existing python library py-spell-checker[15] and (Goldberg and Levy, 2014)'s word2vec word embedding technique. The python spell-check library py-spell-checker[16] checks every word for a misspelling and suggests two or more possible correct words. The original sentence is then parsed through the word2vec (Goldberg and Levy, 2014) model to get obtain its word embedding. The candidate words suggested by the spell check library are then replaced in the original sentence, and the new sentence is parsed through the word2vec model again. We then calculate the cosine distances between the original and possible replacement sentences, and the sentence with the highest cosine score or cosine score above 0.9 i.e., most similar to the original sentence, replaces the original sentence in the dataset.

[15]https://pypi.org/project/pyspellchecker/

[16]https://pypi.org/project/pyspellchecker/

Algorithm 1 Algorithm for contextual misspelled word correction using Word2Vec

import spellcheck() ▷ Python Package
import slang word dictionary ▷ Python dictionary of slang words
import word2Vec ▷ Word2Vec Model
for $sentence$ in $list\ sentences$ **do**
    $spellcheck \leftarrow sentence$
    $word2vec \leftarrow sentence$
    $wordoptions \leftarrow spellcheck(sentence)$
    $new_{sentence}$ ←word options+$sentence$
    $new_{word2vec}$ ←new sentence
    similarity = word2vec.cosine - new$_{word2vec}$.$cosine$
    if similarity > threshold (0.90) then sentence = new$_{sentence}$ −
        return $sentences$

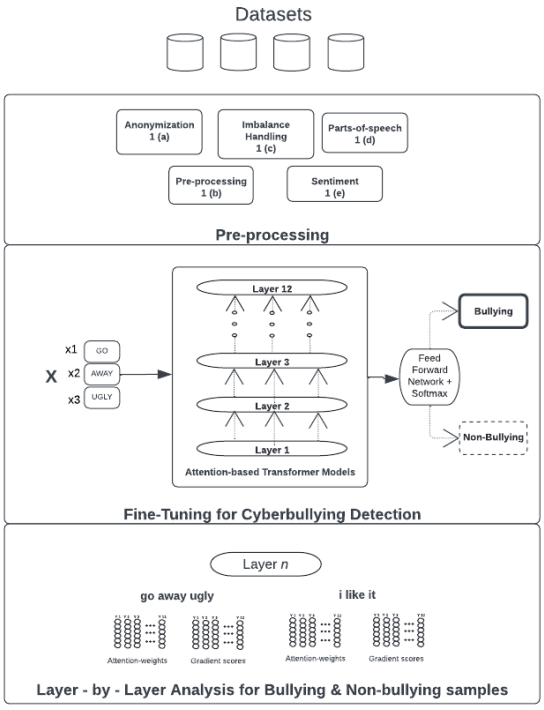

Figure 7: Experiment Schema

A.2 Dataset Split

The table 4 represents each grouped dataset's training, validation, and test-set size.

A.3 Experiment Schema

A.4 Imbalance Handling Results

Table 5 presents the results of no-sampling and over-sampling techniques leveraged in this study

Dataset-Type	Total Size	Training-set	Validation-set	Test-set	
				90% for Performance	10% for Attention
Question-Answering	139,500	111,600	13,950	12,555	1,395
User-Comment	110,323	88,259	11,032	9,929	1,103
Twitter	13,974	11,179	1,397	1,258	140
Gaming	34,229	27,383	3,423	3,081	342

Table 4: Dataset split-size

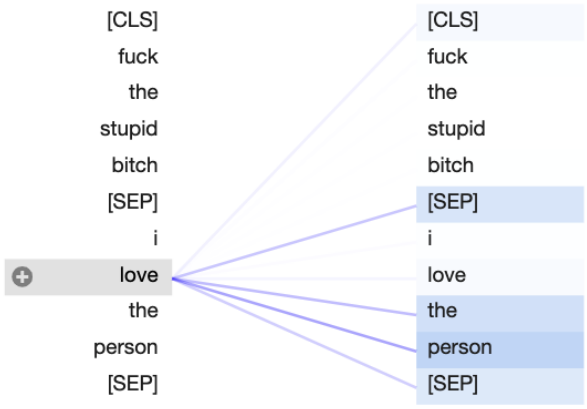

Figure 8: Attention analysis for Positive Sentiment Word

for fine-tuning pre-trained LMs.

A.5 Sentiment Attention Analysis

To analyse if negative and positive sentiment words have more attention in a text sequence. We selected the following two sentences, *"f*ck the st*pid b*tch"* (negative sentiment), and *"i love this person"* (positive sentiment), and visualized the sequence to sequence attention by leveraging BertViz (Vig, 2019b). Our findings for both positive and negative words in sample sentences for almost all layers and attention-heads on both fine-tuned BERT and HateBERT are depicted in Figure 8 and 9. As seen in Figure8, the word *"love"* in positive sentence *"i love the person"*, has higher attention distribution with neutral words like *"the"* and *"person"*. As seen in Figure9, the word *"b*tch"* in negative sentence has higher attention distribution with neutral words like *"i, the"* and *"person"*. This too disproves our hypothesis, and shows attention is not fully at negative sentiment words, instead its similar for positive sentiment words and at times higher for neutral sentiment words.

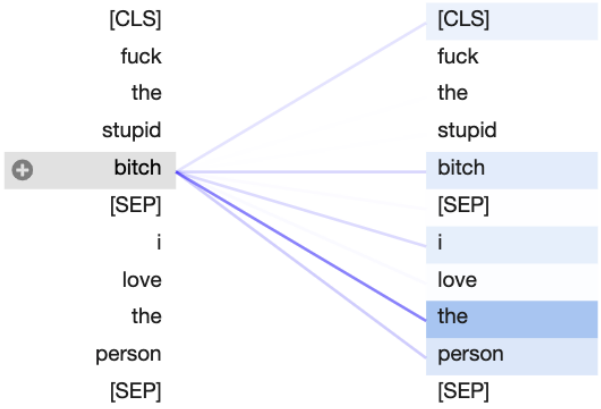

Figure 9: Attention analysis for Negative Sentiment Word

Model	Dataset	No-sampling	Oversampling
BERT+	Gaming	0.7478|	**0.7746**
	UC	**0.8476**	0.8224|
	QA	**0.9496**	0.939|
	Twitter	**0.94**	0.365|
HateBERT+	Gaming	0.6857|	**0.7468**
	UC	**0.8497**	0.8261|
	QA	**0.9498**	0.9076|
	Twitter	**0.939**	0.938|

Table 5: Model's performance on balanced and imbalanced training datasets
Note: UC → user-comment; QA → question-answering