Attention-Weighted Integrated Gradients for Target-Aware Cyberbullying Detection

Sharan Ramjee
sramjee@stanford.edu

Sarthak Consul
sarthakc@stanford.edu

I. Introduction

Cyberbullying detection is a challenging task to tackle, given the complex nature of the problem and the lack of NLP literature addressing this issue. Typical Sentiment Analysis models are susceptible to robustness issues [1] where attacks can be generated by appending positive-sentiment text to negative-sentiment (cyberbullying) text. For instance, the sentence "@SConsul is a terrible person and should be imprisoned. Today is a beautiful day and the weather is amazing." will escape being classified as cyberbullying because while the sentiment of the first part of the sentence negative, the model will classify the overall sentiment of the sentence as positive due to the overwhelming positive sentiment of the second part of the sentence. In order to tackle such issues, we propose Attention-Weighted Integrated Gradients (AWIG) for Target-Aware Cyberbullying Detection using the `twitter-roBERTa-base-sentiment-latest` model [2][3], where the sentiment of the sentence with respect to an aspect-target token ("@SConsul" here) is computed for improved performance. The code is available on GitHub: https://github.com/sharanramjee/cyberbullying-awig.

II. Dataset

The hatespeech-twitter dataset [4] consists of ∼100k tweets with the following labels: "Normal", "Abusive", "Spam", "Hateful". For pre-processing, we combine the tweets from the "Abusive" and "Hate" classes into a single class: "Cyberbullying". Furthermore, we discard tweets belonging to the "Spam" class (probably generated by a bot), leaving us with ∼86k tweets in the dataset (63% "Normal" and 37% "Cyberbullying"). Additionally, we remove the emojis and URLs from all tweets and replace the HTML codes with the characters they represent (`&amp;` = &, `&gt;` = >, `&lt;` = <).

Target-aware Sentiment Analysis requires a target word (also known as the aspect-target) in the sentence. Here, references to `@` usernames are treated as the aspect-targets. There are ∼48k tweets that contain `@`s, out of which, ∼4.3k tweets are randomly sampled. Finally, during our analyses, we notice that the username tokens themselves influence the sentiment of the model, which is undesirable from a fairness standpoint given that a user's username should not have an impact on the model's output. As such, we replace these usernames with a neutral token: `username`. An extensive fairness analysis of the impact of usernames on the model's performance is given in the fairness analysis section.

III. Technical Approach

Our approach to cyberbullying detection leverages Integrated Gradients (IG) attribution [5]. The IG completeness property ensures that the token attributions sum to the output logits of the model. As examined earlier, these token attributions can be misleading as not all parts of the tweet refer to the aspect-target. Our proposed method re-weights the attributions of the tokens with the aspect-target self-attention weights.

*a) **Integrated Gradients (IG):*** A lambda QoI function of the form $\max(positive\_logit, neutral\_logit) - negative\_logit$ is used to compute the IG attributions. The IG baseline is generated by replacing all non-username/standard (i.e. `eos`, `bos`, `cls`) tokens with the `pad` token.

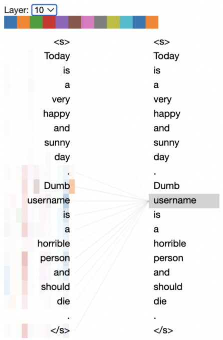

Fig. 1: BertViz [6] visualization of the model self-attention weights w.r.t the target token at layer 10

*b) **Attention Weights:*** Figure 1 illustrates the self-attention weights of the model. It is evident that tokens related to (in this case: words describing) the target word have high attention weights in at least one head of at least one layer. We thus compute a weight factor for every token as:

$$\alpha_i \begin{cases} = 0 & \text{if } token_i = token_t \\ \propto \max_l \max_h \mathrm{A}(l, h, token_i, token_t) & \text{otherwise} \end{cases}$$

where $\mathrm{A}(l, h, token_i, token_t)$ is the attention weight of the model in the l^{th} layer and h^{th} head of $token_i$ towards the target token $token_t$ and $\sum_i \alpha_i = 1$ (weight factors are normalized).

*c) **Attention-weighted Integrated Gradients (AWIG):*** The score of a tweet is computed as the convex combination of the token attributes:

$$\text{AWIG Score} = \sum_i \alpha_i \times \mathrm{QoI}(token_i) \tag{1}$$

A tweet is classified as "Cyberbullying" when its AWIG Score is less than or equal to 0 and as "Normal" otherwise.

IV. Analysis

For each of our analyses, we use the following evaluation metrics: accuracy, precision, recall, and F1-score. For cyberbullying detection, it is more important to maximize true-positives (true-cyberbullying) and minimize false-negatives (false-non-cyberbullying) because there is a higher cost of real cyberbullies escaping being flagged as cyberbullies in comparison to non-cyberbullies being flagged as cyberbullies [7]. That being said, we will pay particular attention to the recall score achieved by the methods during our analyses.

A. Robustness

*a) **Camouflage Attacks**:* As seen in the example in the Introduction section, typical Sentiment Analysis methods fail in cases where a sentence of one sentiment can be appended to another sentence of opposite sentiment to shift the sentiment of the overall sentence in a particular direction of interest [1]. We refer to such attacks as "camouflage attacks" since the sentiment of a particular part of the tweet is camouflaged by the opposing sentiment of other parts of the tweet.

In order to analyze the robustness of our approach to camouflage attacks, we generate an adversarial dataset with 1000 examples (500 positive and 500 negative). The dataset is generated by prepending and appending randomly selected tweets with two randomly selected tweets of the opposing sentiment. The results of the camouflage attack analysis are given in Table I.

TABLE I: Camouflage Attack Analysis

Method	Accuracy	Precision	Recall	F1-Score
Baseline	0.036	0.057	0.060	0.059
AWIG	0.475	0.486	0.848	0.618

We observe that AWIG greatly outperforms the baseline across all the metrics considered. On visualizing the attention-weighted attributions (Figures 4, 5), it is apparent that the self-attention appropriately re-weighs the attribution of the target sentences higher than the neighboring adversarial sentences.

*b) **Perturbation Attacks**:* We analyze the robustness of our methods with respect to token perturbation attacks by generating a dataset of 100 adversarial examples using the following methods in the TextAttack[8] library:

(i) TextFooler [9]: Replace important words in the input tweets with synonyms (semantically similar words) that are extracted using the cosine similarity of the *counter-fitting* word embeddings [10].

(ii) BertAttack [11]: Replace important words in the input tweet with a word suggested by a pretrained BERT model.

(iii) DeepWordBug (DWB) [12]: Mis-spell important words in the input tweet.

Given that all three attacks are black-box attacks (no access to gradients), the importances of words are estimated by ranking the differences (larger difference corresponds to higher importance) of the output scores upon removing (or replacing with the `[MASK]` token) each of the words. The results of the perturbation attack analysis is given as performance drops (lower is better) resulting from such attacks, given in Table II.

TABLE II: Performance Drops on Perturbation Attacks

Attack	Model	Accuracy	Precision	Recall	F1-Score
TextFooler	Base Model	0.825	1.0	1.0	1.0
	AWIG	0.285	0.092	0.361	0.259
BertAttack	Base Model	0.803	1.0	1.0	1.0
	AWIG	0.264	0.137	0.284	0.222
DWB	Base Model	0.707	0.563	0.741	0.674
	AWIG	0.408	0.277	0.391	0.343

We observe that AWIG is less vulnerable compared to the baseline across all the perturbation attacks considered. This increased robustness is due to the fact that the AWIG attention-weights reduce the attributions of highly important tokens that are irrelevant to the aspect-target. It is worth noting that AWIG does not guarantee an increase in robustness as the perturbed tokens do not necessarily have low attention weights.

B. Fairness

*a) **Twitter Usernames**:* In many cases, we noticed that the Twitter usernames impact the output of the model. This is undesirable because usernames can sometimes reflect users' protected attributes [13]. We measure this impact by extracting all the `@<username>` aspect-targets from our dataset and computing their sentiments using AWIG. The most negative and positive tweets found (as measured by output logits) were `@DumbPeopleAsf` and `@BestVinesEver`, respectively. We then create a positively and negatively biased datasets by replacing all the usernames in the hatespeech-twitter dataset by `@DumbPeopleAsf` and `@BestVinesEver`, respectively. Finally, we run inference on these biased datasets using AWIG and plot histograms of the positive, neutral, and negative output logits, given in Figure 2.

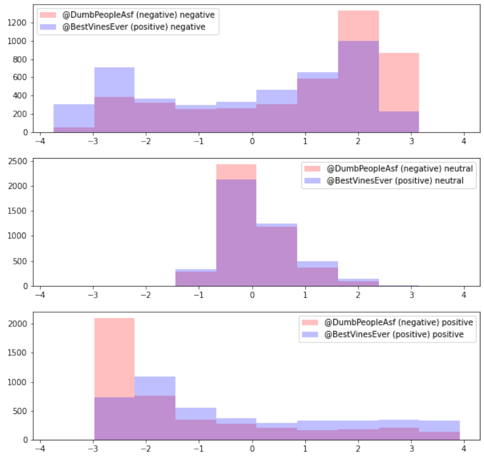

Fig. 2: Histogram of output model logits on the biased datasets

Given the skews in the histograms, we observe that indeed, usernames drastically impact the score of the model. We delve deeper into the analysis by extracting the top three tweets with the highest differences in magnitudes among the output logits using the two biased usernames and visualize their IG attributions, given in Figure 3 of Appendix B. As mentioned in the Dataset section, we mitigate the effect of the usernames on the model outcome by pre-processing the dataset to replace all usernames with a neutral word: `username`. The performance of the methods on the non pre-processed and pre-processed datasets is given in Table III. We observe that using a neutral username improves performance across all the metrics considered. It is impossible to completely eliminate the influence of usernames using conventional methods of Sentiment Analysis. However, we make use of an IG baseline in AWIG that incorporates the `@username` token into the attributions to completely eliminate its effect on the prediction outcome (`@username` is given an attribution of 0). Therefore, by virtue of AWIG's prediction mechanism, we achieve fairness by assessing the sentiment of tweets without taking usernames into account.

TABLE III: Twitter Username Analysis

Method	Pre-processed	Accuracy	Precision	Recall	F1-Score
Baseline	No	0.789	0.735	0.865	0.795
Baseline	Yes	0.795	0.738	0.881	0.803
AWIG	No	0.799	0.743	0.886	0.805
AWIG	Yes	0.805	0.746	0.892	0.813

*b) **Protected Attributes**:* We explore the fairness of the methods on the following protected attributes: race, sex, and political leaning. Given that the hatespeech-twitter dataset lacks these attributes, we use external models to predict each of the them for every tweet in the dataset prior to computing the following fairness metrics: demographic parity difference/ratio, equalized odds difference/ratio, false negative/positive rates, and true negative/positive rates (equality of opportunity). Each of these external models comprise of XGBoost Boosted Decision Trees [14] trained using a bag-of-words feature encoder [15] with 5000 features. It is important to note that these external models were trained on other Twitter datasets where these attributes were available and did not achieve perfect performance on any of them (model accuracies can be found in their respective papers). That being said, the protected attribute labels predicted by these models on the hatespeech-twitter dataset are noisy and as such, the fairness metrics should be taken with a grain of salt.

(i) **Race**: The race classifier is trained on a Twitter corpus [16] with the labels: 0 for `African-American English` and 1 for `White-Aligned English`. The fairness metrics for these groups on the hatespeech-twitter dataset is given in Table IV. We observe that while the baseline outperforms AWIG in terms of demographic parity and equalized odds, AWIG outperforms the baseline in terms of equality of opportunity. Both the baseline and AWIG models tended to classify AAE tweets as cyberbullying more than than WAE as they lack social context [17] [18] and end up discriminating against the minority group. For instance, it would be considered okay for a black person to use the N-word but not a white person.

TABLE IV: Race Fairness Analysis

Metric	Baseline	AWIG
Demographic Parity Difference	0.285	0.334
Demographic Parity Ratio	0.641	0.599
Equalized Odds Difference	0.041	0.092
Equalized Odds Ratio	0.873	0.745
False Negative Rate	0.119	0.108
False Positive Rate	0.282	0.274
True Negative Rate	0.718	0.726
True Positive Rate	0.881	0.892

(ii) **Sex**: The sex classifier is trained on a Twitter corpus [19] with the labels: 0 for `female` and 1 for `male`. The fairness metrics for these groups on the hatespeech-twitter dataset is given in Table V. We observe that AWIG outperforms the baseline across all the fairness metrics.

TABLE V: Sex Fairness Analysis

Metric	Baseline	AWIG
Demographic Parity Difference	0.012	0.012
Demographic Parity Ratio	0.979	0.980
Equalized Odds Difference	0.119	0.109
Equalized Odds Ratio	0.881	0.891
False Negative Rate	0.119	0.108
False Positive Rate	0.282	0.274
True Negative Rate	0.718	0.726
True Positive Rate	0.881	0.892

(iii) **Political Leaning**: The political leaning classifier is trained on a Twitter corpus [20] with the labels: 0 for `republican` and 1 for `democrat`. The fairness metrics for these groups on the hatespeech-twitter dataset is given in Table VI. We observe that AWIG outperforms the baseline across all the fairness metrics considered.

TABLE VI: Political Leaning Fairness Analysis

Metric	Baseline	AWIG
Demographic Parity Difference	0.297	0.279
Demographic Parity Ratio	0.485	0.517
Equalized Odds Difference	0.156	0.115
Equalized Odds Ratio	0.464	0.589
False Negative Rate	0.119	0.108
False Positive Rate	0.282	0.274
True Negative Rate	0.718	0.726
True Positive Rate	0.881	0.892

V. Conclusions

In this paper, we propose a new method for target-aware cyberbullying detection on the hatespeech-twitter dataset: Attention-Weighted Integrated Gradients (AWIG). AWIG is modular and can be applied to any Sentiment Analysis transformer model (with self-attention) to make the system more target-aware. Through our robustness analyses, we show that AWIG outperforms the baseline when dealing with camouflage and perturbation attacks. Furthermore, through our fairness analyses, we observe and neutralize the impact of Twitter usernames on the prediction outcome. Finally, we show that AWIG outperforms the baseline in the context of fairness across the protected attributes considered (race, sex, political leaning).

References

[1] X. Xing, Z. Jin, D. Jin, B. Wang, Q. Zhang, and X. Huang, "Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis," *arXiv preprint arXiv:2009.07964*, 2020.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[3] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, "Timelms: Diachronic language models from twitter," *ArXiv*, vol. abs/2202.03829, 2022.

[4] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *11th International Conference on Web and Social Media, ICWSM 2018*, AAAI Press, 2018.

[5] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*, pp. 3319–3328, PMLR, 2017.

[6] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Florence, Italy), pp. 37–42, Association for Computational Linguistics, July 2019.

[7] C. Chen, S. Ramjee, and J. Wang, "Aspect-target sentiment classification for cyberbullying detection."

[8] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi, "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 119–126, 2020.

[9] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? a strong baseline for natural language attack on text classification and entailment," in *AAAI*, 2020.

[10] N. Mrksic, D. Ó. Séaghdha, B. Thomson, M. Gaić, L. M. Rojas-Barahona, P. hao Su, D. Vandyke, T.-H. Wen, and S. J. Young, "Counter-fitting word vectors to linguistic constraints," in *NAACL*, 2016.

[11] L. Li, R. Ma, Q. Guo, X. Xue, and X. Qiu, "Bert-attack: Adversarial attack against bert using bert," *ArXiv*, vol. abs/2004.09984, 2020.

[12] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56, 2018.

[13] B. Erşahin, Ö. Aktaş, D. Kılınç, and C. Akyol, "Twitter fake account detection," in *2017 International Conference on Computer Science and Engineering (UBMK)*, pp. 388–392, IEEE, 2017.

[14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[15] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1, pp. 43–52, 2010.

[16] S. L. Blodgett, L. Green, and B. T. O'Connor, "Demographic dialectal variation in social media: A case study of african-american english," in *EMNLP*, 2016.

[17] M. Sap, D. Card, S. Gabriel, Y. Choi, and A. N. Smith, "The risk of racial bias in hate speech detection," in *ACL*, 2019.

[18] T. Davidson, D. Bhattacharya, and I. Weber, "Racial bias in hate speech and abusive language detection datasets," *arXiv preprint arXiv:1905.12516*, 2019.

[19] F. Eight, "Twitter user gender classification," in *Kaggle*, 2017.

[20] K. Pastor, "Democrat vs. republican tweets," in *Kaggle*, 2018.

Appendix A
Camouflage Attack Analysis

In order to delve deeper into how each of the models deal with such camouflage attacks, we plot the token attributions of the baseline and AWIG for two such examples, given in figures 4 (positive sentiment) and 5 (negative sentiment) in the next page. We notice that the baseline fails to correctly classify both examples because the opposing attributions of the preceding and succeeding sentences camouflage the attributions of the target sentence. However, AWIG is able to correctly classify both examples because the attention weights allow the attributions of the target sentence to weighted higher and the attributions of the preceding and succeeding sentences to be weighted lower, thus leading to improved performance.

Appendix B
Twitter Username Analysis

In order to delve deeper into how usernames impact the model outputs, we extract the top three tweets with the highest difference in magnitude among the baseline output logits using the two biased usernames and visualize their Integrated Gradients attributions, given in Figure 3. We observe that in such cases, the usernames can impact the outputs of the models to such an extent that the sentiments of the tweets are misclassified.

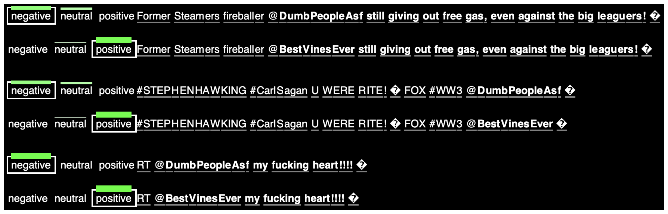

Fig. 3: Integrated Gradients attributions of the top three tweets with the highest difference in magnitude among the baseline output logits using the two biased usernames

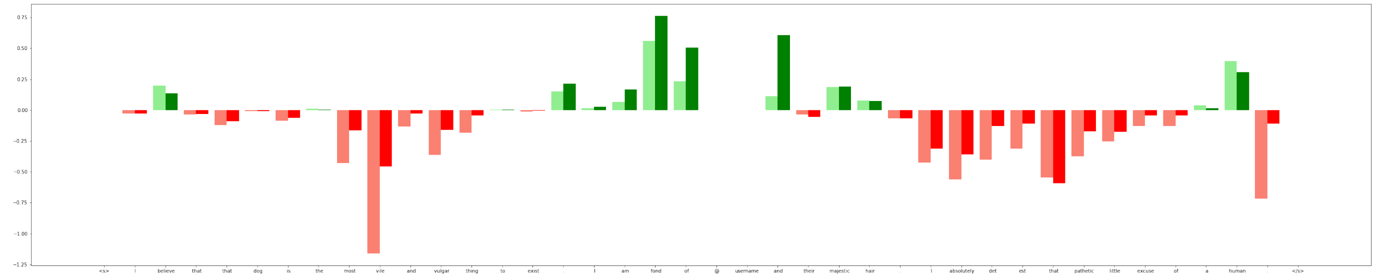

Fig. 4: Camouflage attack analysis of a positive example for the baseline (light green/light red) and AWIG (green/red).

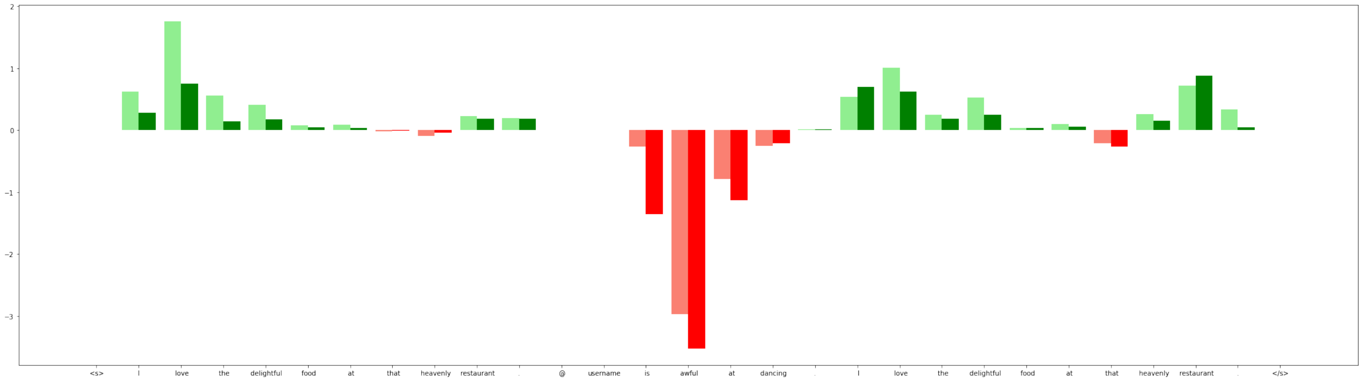

Fig. 5: Camouflage attack analysis of a negative example for the baseline (light green/light red) and AWIG (green/red).