Transfer Learning and Lexicon-Based Approaches for Implicit Hate Speech Detection: A Comparative Study of Human and GPT-4 Annotation

Saad Almohaimeed
Dept. of Computer Science
University of Central Florida
Orlando, USA
sa583575@ucf.edu

Saleh Almohaimeed
Dept. of Computer Science
University of Central Florida
Orlando, USA
sa247216@ucf.edu

Ladislau Bölöni
Dept. of Computer Science
University of Central Florida
Orlando, USA
ladislau.boloni@ucf.edu

***Abstract*—Detecting harmful speech is the subject of significant research effort both in the academia and industry. While good progress was made on detecting explicit hate speech, detecting implicit hate remains difficult as it requires a deep understanding of the allusions of the text and the social context in which it was uttered. In this paper we study the effectiveness of several approaches to implicit hate speech detection, including lexicon-based approaches, transfer learning, and the use of up-to-date large language models, such as GPT-4. By combining lexicon-based approach with the targeted topics, we performed transfer learning experiments using knowledge from seven public harmful speech datasets. Various combinations of the proposed approaches showed an improvement of 0.6 - 2.3% in the F1-Macro score compared to the baselines. We observed that while GPT-4 annotations show a good agreement with human labels, there is often a conflict when interpreting sarcasm, text shortening based on context, and speech that targets individuals.**

Warning: due to the nature of the research subject, this paper contains explicit and potentially offensive language.

***Index Terms*—implicit hate, hate speech, abusive language, transfer learning, lexicon-based, GPT-4 annotation**

I. Introduction

In recent years, significant research has been conducted on topics associated with hate speech such as offensive language, cyberbullying, abusive language or toxic content. Several projects addressed the problem in general [1], [2] while others examined more specific forms of hate speech, including sexism and racism [3], white supremacism [4], and hate speech in sports [5]. Other projects targeted topics beyond the detection of hate speech including recognizing the targeted group [6]–[8], explainable hate speech [9] and cross-lingual and cross-domain hate speech detection [8], [10], [11].

One of the hardest current challenges of this research area is to detect *implicit hate speech*. Such texts do not contain overt profanity or slurs, but at the semantic level contain references to topics that imply hatred. Automatic detection of such implicit hate speech is made difficult by the need of external information beyond the text itself. The meaning of implicit hate speech might even be concealed from neutral human audiences who are not familiar with the specific references.

In this paper, we study a number of approaches for the classification of a given text as implicit hate speech (IH) or not implicit hate speech (NIH). Our first technique uses a *lexicon* to aid the machine in comprehending the given text better than raw text alone. Two separate lexicons have been built for covering slang and offensive language respectively. Afterward, we introduce a transfer learning approach that utilizes previously published harmful speech datasets to improve the detection of implicit hate speech. The final part of our study compares GPT-4's effectiveness in detecting implicit hate speech with that of human judgment. To conduct our experiments, we utilized the THOS dataset [6], our previous work that considered implicit hate speech and targeted topics (TPCs) as part of our proposed features.

The main contributions of this paper are as follows:

- Unified and aligned the features of seven public harmful speech datasets. Also, we have compiled and crawled eight different sources to create two lexicons of slang and offensive words.
- Transfered the knowledge of the unified datasets to improve implicit hate speech detection.
- Performed a comprehensive set of experiments using transfer learning, lexicon-based approaches, targeted topics and GPT-4 annotations.

II. Related Work

A. Transfer Learning for Hate Speech

Transfer learning has proven effective in various fields, notably in hate speech detection [11] [12]. Pamungkas and Patti [11] demonstrated its efficacy by training source models on diverse harmful speech datasets in multiple languages, incorporating the HurtLex lexicon. They utilized LSVC and LSTM for both source and target models, observing that lexicons enhance performance. Their findings highlighted transfer learning's robustness across domains but noted challenges when adapting to different languages.

Mozafari et al. [12] enhanced hate speech detection using transfer learning with Wikipedia and BookCorpus, applied to datasets [1] and [3]. Their approach, incorporating BERT's fine-tuning layers, significantly improved embedded text information, leading to better detection outcomes.

B. Lexicon-Based Approach

Lexicon-based methods are prevalent in domain-specific tasks due to their domain independence, allowing usage across various platforms like Twitter and Facebook [13]. However, employing an offensive lexicon for hate speech detection can lead to false positives, as offensive words may appear in non-hateful contexts [14]. Additionally, human annotators might incorrectly label data when relying solely on offensive lexicons, overlooking semantic context [1]. Therefore, combining offensive lexicons with annotations can aid models in differentiating between hate and merely offensive language.

According to Vashistha and Zubiaga [10], the brevity and slang in social media text hinder model performance. Alatawi et al. [4] assessed deep learning's effectiveness on hate speech detection, using a 2k dataset annotated via crowd-sourcing with a lexicon of slangs and hate speech terms. They compared BiLSTM with lexicon embedding to a pretrained BERT model. While BERT outperformed BiLSTM, it struggled with slang and coded language crucial for interpreting hate speech. Thus, incorporating a slang lexicon is expected to enhance performance in domain-specific tasks.

C. Implicit Hate

ElSherief et al. [15] created a dataset with 21k tweets, encompassing diverse implicit hate speech types. They analyzed it using SVM (with TF-IDFs, n-grams, GloVe) and BERT, exploring features like data augmentation, knowledge graph and multi-label classification. The authors found that BERT surpassed SVM in performance, but knowledge graphs didn't enhance implicit hate detection. The models struggled with coded symbols in texts, like "#NationalSocialism" and "#WPWW".

D. GPT-4

Huang et al. [16] examined GPT-4's performance in identifying implicit hate speech. Comparing its annotations to human ones using the dataset from [15] they reported that GPT-4 correctly classified 80% of instances. Further analysis with crowd-sourced verification showed a higher agreement with GPT-4's decisions than human annotations. The study highlighted GPT-4's effectiveness in providing detailed Natural Language Explanations (NLE), enhancing decision accuracy.

III. Methodology

Fig. 1 illustrates our methodology for comparing the efficacy of different approaches and features by dividing them into separate tasks and subtasks.

A. Flagger

The Flagger (source model) is a RoBERTa [17] model that has been fine-tuned using seven public datasets of harmful speech such as hate speech, abusive, cyberbullying, and offensive language. Since different datasets have different labels (e.g. offensive, hate, racism, sexism, abusive, disrespectful, threats), we have unified all datasets into two labels (normal, flagged). Thus, flagged speech is any speech that has been identified as "not normal" in the seven datasets. We believe that identifying a general view of the data will simplify the task for the next model (i.e. target model) to narrow down to the correct answer. Therefore, the output of the Flagger model for any suspicious text will be as follows: $original\_text\ [SEP]\ flagged.$

Datasets: in order to identify any suspicious text, we intend to train our model with an extensive dataset. So, to obtain a large corpus of carefully annotated data, we utilized a set of benchmark datasets [1]–[3], [7]–[9], [18] published in the field of hate speech and harmful languages, as shown in Table I. These datasets resulted in a total of 134k rows.

Filtering Datasets: it has been noted that some papers release datasets without the actual text (e.g. only post_id), which requires us to retrieve the corresponding text from the source platform. In most cases, harmful or offensive content on social media platforms is removed by the author or the platform if the post violates its policies and guidelines. This removal has resulted in a reduction in the size of the dataset in comparison with its original size. As well, some datasets were shortened from data which was not considered in this study (e.g., non-English text was removed from Ousidhoum et al.'s dataset [8]). In the Founta et al. [2] dataset, we considered the *Spam* label as normal speech since it does not convey harmful information. In addition, for Gautam et al. dataset [18], only the 'hate speech' label is flagged while other labels are considered normal. Regarding datasets that provide more than one annotator decision [9], the majority vote was considered. In case of voting equality, we considered them as 'flagged' label.

Source Model Bias: LLMs such as RoBERTa are generally easy to train on sentiment datasets in order to differentiate between positive and negative sentiments (e.g. normal speech and offensive speech). A significant portion of our combined datasets is comprised of offensive raws. It is likely that the model that will be trained on this dataset will be biased in favor of raws that contain explicit slurs or offensive language. In this study, however, our purpose is to detect harmful speech that is free of explicitnesses. The detection of such suspicious text will help the target model narrow down to the IH. Accordingly, we removed any row from the combined dataset that contains offensive language using the offensive-lexicon we will introduce in section III-B. Therefore, the final dataset size became 97k. This is the dataset that was used in order to train, validate, and test the Flagger.

B. Lexicon-Based Interpreter

Slang Interpreter (SI): is a tool that changes the corresponding slang, coded language or abbreviation to its interpretation.

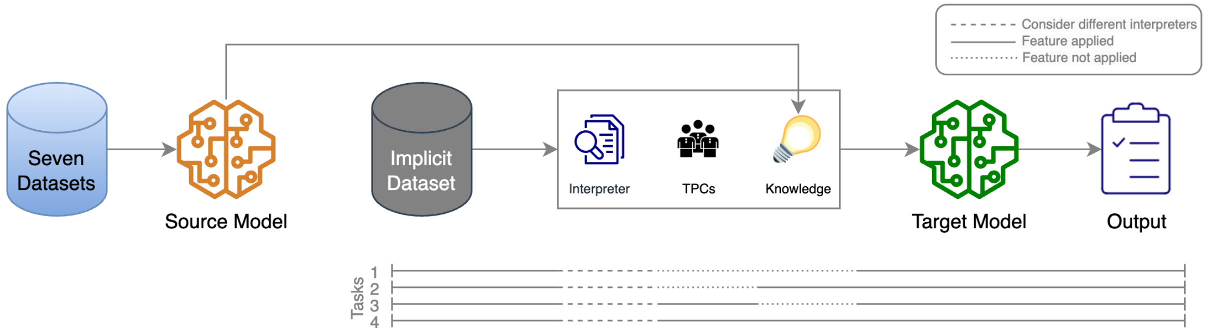

Fig. 1: The pipeline of our proposed methodology

TABLE I: Combined Datasets

Name	Labels	Size	Normal	Flagged
Waseem and Hovy [3]	sexism, racism, neither	16907	10423	6484
Davidson et al. [1]	hate, offensive, normal	24783	2641	22142
Founta et al. [2]	offensive, abusive, hateful speech, aggressive, cyberbullying, spam, normal	45982	38952	7030
(OLID) Zampieri et al. [7]	offensive, not offensive	13240	8382	4858
Ousidhoum et al. [8]	Abusive, Hateful, Offensive, Disrespectful, Fearful, Normal	5647	371	5276
(MeTooMA) Gautam et al. [18]	relevance, stance, hate speech, sarcasm, dialogue acts	7814	6993	821
(hateXplain) Mathew et al. [9]	hate, offensive, normal	20148	4606	15542
Total		134521	72368	62153

We have built a slang-lexicon that contains a list of slangs, coded language or abbreviations with their meanings or long format. In order to improve the model's understanding of the annotations, a clear text must be provided as we aimed from building this tool. As an example, here is an abbreviation and coded slang from Founta's dataset [2]:

1) *You are doing a fine job being an **a$$hole w/o** their help!*
2) *You are doing a fine job being an **a**hole**[1] **without** their help!*

In the second example, a coded slang expression *a$$hole* is interpreted to its actual form. Also the word *without* is used more frequently than its abbreviation *w/o*. The use of this interpretation can assist the model encoding in understanding the meaning instead of using ambiguous abbreviations or coded language. Slang-lexicon was crawled from two sources[2][3]. A total of 6k slangs, abbreviations, and coded languages were crawled. There are some slangs that are simply ordinary common words that can only be understood in their context. For example, the word *ice* is typically used to refer to frozen water, but it is also used to refer to a special type of drug. A further example would be *sick* which is a word that describes illness; however, it is also a slang word that means *cool, great or excellent*. For this reason, we manually filtered the crawled slangs to remove any common slang that is only interpretable by their context. The lexicon available in Github[4], including 4.8k slang terms and their meaning.

Offensive Interpreter (OI): is a tool designed to determine whether the given text contains slurs or offensive words. We have built a lexicon that contains a list of offensive words. The objective is to provide the model with a more detailed view of the text content. The following is an example of offensive interpretation from Founta's dataset [2]:

1) *You are doing a fine job being an a**hole without their help!*
2) *You are doing a fine job being an a**hole without their help! **[SEP] slurs***

In the second example, the text contains more information regarding the offensiveness of the given text (i.e. slurs). The word *a**hole* is a commonly used offensive word in social media that may be understandable by some pre-trained models. However, less common offensive words cannot be captured by the pre-trained model (e.g. *of*y, eun*ch, bitc**yke*). Accordingly, having a more in-depth interpretation of the text, as in the second example, will provide a greater amount of information. The offensive-lexicon was collected and crawled from Wikipedia religious slurs[5], Wikipedia ethnic slurs[6], Wikipedia slurs list[7], Google profanity list[8], Noswearing.com[9]. We also collected 1.3k offensive words provided by Carnegie Mellon University[10]. There are,

[1] Offensive terms were obscured with asterisks (*) in this paper.
[2] https://slang.net/
[3] https://www.berlitz.com/blog/american-slang-words
[4] https://github.com/mohaimeed/offensive-slangs-lexicons
[5] https://en.wikipedia.org/wiki/List_of_religious_slurs
[6] https://en.wikipedia.org/wiki/List_of_ethnic_slurs
[7] https://en.wikipedia.org/wiki/List_of_ethnic_slurs_and_epithets_by_ethnicity
[8] https://github.com/coffee-and-fun/google-profanity-words/blob/main/data/list.txt
[9] https://www.noswearing.com/dictionary
[10] https://www.cs.cmu.edu/ biglou/resources/bad-words.txt

TABLE II: THOS dataset distribution for implicit hate speech annotation by Human and GPT4

Annotators	Speech Type	number of samples
Human	NIH	6404
	IH	1878
GPT4	NIH	5700
	IH	2582
Total for Each		8282

however, many words that represent a normal set of words (e.g. White, Black, Arab, American, Muslim, Christian, ..etc). Therefore, we did not include this list in our experiments. The offensive-lexicon available in Github[11], consisting of 1.8k term.

C. Implicit Hate Detection

Implicit Hate Detection is our **target model** in the transfer-learning process. Our objective is to improve the implicit hate speech detection by using the THOS dataset [6], which has been annotated including implicit hate speech as a defined label (i.e. hs). THOS has examined various aspects of hate speech and offensive language (e.g. explicit hate, implicit hate, targeted topics, and subtopics). We considered only implicit hate speech and targeted topics (TPCs) in our study. As part of our experiment, we considered SI, OI, targeted TPCs and transferred the knowledge gained from Flagger to the aforementioned approaches and features. Experiment details and results discussed in IV and V.

D. GPT-4

1) Annotation: we used ChatGPT/GPT-4 API to annotate THOS dataset into categories: normal, explicit, and implicit hate speech, with explicit hate being outside our scope. We observed GPT-4's improved accuracy in classifying implicit hate using a three-category option (normal, explicit hate, or implicit hate) over a binary one. The annotation prompt was: "Please classify the given text as (explicit hate, implicit hate, normal speech) as a single token response: $given\_sample$". Despite requesting single-token responses, GPT-4 provided more detailed answers, demonstrating its understanding. Annotations were collected on September 14th, 2023, and compared with human annotations in Table II.

2) Evaluation: In order to verify the accuracy of GPT-4's decision on the conflicts between human and GPT-4 annotations in Table III, we conducted a manual evaluation by two experts in the field of hate speech. For the verification of the experts' decisions, we selected a random 200 samples(100 where human chose IH and 100 where GPT-4 choose IH) and applied Cohen's Kappa, the inter-annotator agreement (IAA) method for the experts' decisions.

TABLE III: THOS dataset human vs. GPT4 annotation on both viewpoints for NIH and IH label

Annotations by	GPT-4 **IH**	GPT-4 **NIH**	
Human **IH**	855	1023	-
Human **NIH**	1727	4677	-
Total matches between human and GPT-4			5532
Total conflicts between human and GPT-4			2750

IV. Experiments Settings

A. Preprocessing

As a preprocessing step prior to getting the text into the model for training, we performed a number of different steps. As a first step, we lowered the text's case since the SI and OI tools are case-sensitive. Next, we removed extra spaces, punctuation marks, URLs, user mentions, and hashtags. Although mentions and hashtags contain valuable information, they usually consist of a concatenated string (e.g. #thursdaymorning, @user337651) that adds additional noise to the text.

B. SI and OI

We have to note that the step of SI and OI are called twice and consecutively. The first is right after lowering case in section IV-A. The second is after applying all preprocessing steps. The reason behind this technique is to capture slang and offensive text that consist of punctuation (e.g. sh!t). Also the consecutive technique of SI and OI is to capture the slangs that are not exist in the offensive-lexicon (e.g. mf) while its slang interpretation exist in the offensive-lexicon.

C. Source Model Settings

The Flagger has been trained, validated, and tested on 58.5k, 19.5k, and 19.5k samples, respectively using RoBERTa-base model. The testing set consisted of 14.4k negative (normal speech) and 5k positive (flagged speech). The model converges in the 4th epoch where the lowest cross-entropy loss is achieved and the highest F1-macro score is reached. For the training and validation sets, the learning rate is 2e-6 and the batch size is 16.

D. Target Model Settings

The target model has been trained, validated and tested on 4.9k, 1.6k, 1.6k samples respectively using RoBERTa-base model. The testing set consisted of 1.3k negative (NIH) and 349 positive (IH) samples. The model converges in the 5th epoch where the lowest cross-entropy loss achieved and highest F1-macro score reached. For the training and validation sets, the learning rate is 2e-6 and the batch size is 16.

V. Results

A. Source Model

Precision and recall were 92% for normal speech as positive classes. In addition, the accuracy and AUROC results were of 88.8% and 92.18% consecutively. However, in the context of this study, we took into account "implicit hate" as a positive class in the precision and recall metrics. Moreover, we have not taken into account accuracy and AUCROC metrics since

[11] https://github.com/mohaimeed/offensive-slangs-lexicons

TABLE IV: Flagger model has been trained, validated and tested as 60%, 20%, 20% consecutively (testing on 19.5k samples)

Model	Precision	Recall	F1-macro
RoBERTa-base	76.1	76.1	83.9

they are less meaningful for unbalanced datasets with positive classes constituting only a quarter of the testing dataset (as explained in IV-C).

As shown in Table IV the Flagger results a competitive precision and recall for the positive class (i.e. Flagged). Finally, the F1-macro achieved 84%, which weighed both the negative and positive harmonic mean of precision and recall equally. The level of results attained at this point indicates that the source model is mature enough to transfer knowledge to the target model.

B. Target Model

The target model got different results showing the competitive scores between different approaches and features. For all the experiments settings, the accuracy results were range from 80% to 83.5%. Also, the AUCROC results were range from 83% to 88%. However, these metrics does not fit on our target model testing dataset as happened with the source model due to unbalanced test dataset (check IV-D).

1) Standalone: on the basis of this baseline model, we will be able to determine how effective the different proposed techniques are in detecting implicit hate speech compared to the standalone model. In this sub-task1, we only preprocessed the text and fed it to the RoBERTa-base model. For precision, recall, and F1-macro, the model achieved 57.1%, 54.8%, and 71.9%, respectively. The baseline results are underlined in Table V.

2) With Flagger: as shown in Table V, Tasks 1 and 2, the Flagger can increase implicit hate speech detection performance by 1.4% in F1-macro score. In addition, Flagger seems to have reduced the number of FN in Task 2 compared to Task 1, as the recall score has increased while maintaining a similar precision score in the first three rows. As a result, F1 performance in txt, txt+SI, and txt+OI improved.

3) SI and OI: for the SI, it appears that it cannot function as a standalone feature within the model. RoBERTa, for instance, performs poorly with SI, leading to more false negative predictions, where instances of implicit hate (IH) are predicted as not IH. Our analysis shows that some common slangs, like *'lol: laugh out loud'* and *'smh: shake my head'*, do not significantly alter the annotated label decision and are less disruptive to the model's predictions than their interpretations. Conversely, the OI approach demonstrates better performance across all tasks but tends to reduce precision by generating more false positives, as noted by Schmidt [14]. Interestingly, combining SI and OI (i.e., SOI) surprisingly yields the best performance across all tasks. The offensive lexicon may overlook coded slangs, but SI interpretations, like *shi+* as *sh*t* and *mf* as *mother f**ker*, are included. These interpretations often lead to correctly identifying explicit hate speech as true negatives, thereby lowering false positives.

4) TPCs: In the field of hate speech, our previous work [6] demonstrated how TPCs can improve the model's performance. Therefore, we included this feature in order to compare it with other proposed techniques for detecting implicit hate speech. As a result, TPCs appear to also improve implicit hate speech detection. The TPCs and Flagger seem to perform better when used separately (Task 2 and 3) than when combined together (Task 4).

C. GPT-4

1) Observation: despite the fact that GPT-4 includes individuals in its definition of hate speech, GPT-4 has been observed not to classify samples as explicit or implicit hate speech when the target of the speech is an individual (e.g. *'indv_name is the puppet master'* or *'indv_name is twatzi'*). Also, GPT-4 appears to be biased towards classifying speech as IH speech when a vulnerable target is mentioned (e.g. *'me (I am Asian and I am forced to eat rice every day)'*). Also, GPT-4 seems to be sometimes confuses on sarcasms to specific group (e.g. *'all X_GROUP looks the same to me'*) where X_GROUP has common look or shape characteristics. Furthermore, opinions of people can be interpreted as implicit hate by GPT-4 (e.g. *'Women make much better music than most of men'*). The GPT-4 explained that its decision was based on its belief that the aforementioned example imply superiority of one gender over another, and therefore the example implies hatred of one gender against another. Additionally, many examples have been observed to be dependent on the author, context, or referenced content where it is difficult to judge by human or GPT-4 from the text itself (e.g. *'We need to vaccinate all of X_GROUP'*, *'She wishes it was Hispanic'*, *'Chinese or some other race'*). Considering the aforementioned examples, we need more information about the author's characteristics and the context of the text in order to determine whether it conveys implicit hate speech. Nevertheless, the GPT-4 was biased toward implied hatred for text that was shortened from more context.

2) Expert Annotators: On the 200 samples annotated by the experts for evaluating the human and GPT-4 annotations, the experts agreed on 169 and disagreed on 31 samples. 59.0% (118 samples) of expert_one annotations were in agreement with human annotations, and 60.5% of expert_two annotations were in agreement with human annotations in THOS. Accordingly, expert_one and expert_two agreed with GPT-4 annotations by 41.0% and 39.5%, respectively. Cohen's kappa agreement was calculated on the two expert annotations and the result was 0.64, which indicates substantial agreement between the two experts.

3) Experiment: In this experiment we used the GPT-4 annotation as a predictive model output and applied metrics

TABLE V: Performance results of Implicit Hate Speech Detection.

The experiment results is the average of 5 different random seeds. Due to imbalance dataset, the F1 score is macro to treat every class as equal. On the other hand, Precision (P) and Recall (R) is not the macro, considering Implicit Hate Classes as positive.

			Task 1			Task 2			Task 3			Task 4		
						With Flagger			With TPCs			Flagger + TPCs		
	Model	Input	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Human Annotation	RoBERTa-base	txt	57.1	54.8	71.9	57.4	59.1	**73.3**	56.6	59.7	73.0	58.2	54.3	72.2
	RoBERTa-base	txt+SI	57.4	53.7	71.7	56.0	58.8	**72.5**	59.3	51.4	71.4	56.7	56.3	72.2
	RoBERTa-base	txt+OI	56.8	59.0	73.1	54.8	64.4	73.5	56.9	63.2	**74.1**	56.4	63.1	73.9
	RoBERTa-base	txt+SOI	57.1	63.3	74.2	56.8	62.8	73.9	57.6	63.5	**74.4**	57.2	63.7	74.1

to it while considering the human annotations as ground-truth labels.

As a result, the GPT-4 model achieved 33.1% precision, 45.5% recall, and 57.8% F1-macro scores. We noted that GPT-4 frequently misclassifies hate speech targeting individuals and often struggles with discerning opinions, sarcasms, and short contexts, leading to suboptimal performance.

VI. Conclusion

In this paper, we introduced a set of approaches and features to detect implicit hate speech. Our findings demonstrate that Flagger can boosts implicit hate speech detection and, when SI combined with OI (as SOI), compensates for SI's limitations as a standalone feature. While TPCs showed promise, they sometimes conflicted with Flagger. In real-life scenario, TPCs require a preliminary model for identification, like Flagger. However, Flagger showed improved performance in scenarios free of lexicons. So, utilizing transfer learning, we achieved comparable outcomes to domain-specific lexicons without needing extensive pre-definition or maintenance. Finally, our study on GPT-4's annotation accuracy revealed its proficiency in correctly identifying implicit hate speech in most instances. However, GPT-4 faced challenges distinguishing sarcasm and opinion, often misclassifying them as implicit hate or normal speech in ambiguous contexts.

Acknowledgements: We would like to express our gratitude to the anonymous reviewers for their insightful feedback. This work was partially supported by the National Science Foundation under award 2114948.

References

[1] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proc. of the Int. AAAI Conf. on Web and Social Media*, vol. 11, no. 1, 2017, pp. 512–515.

[2] A. 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 *Proc. of the Int. AAAI Conf. on Web and Social Media*, vol. 12, no. 1, 2018.

[3] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on Twitter," in *Proc. of the NAACL Student Research Workshop*, June 2016, pp. 88–93.

[4] H. S. Alatawi, A. M. Alhothali, and K. M. Moria, "Detecting white supremacist hate speech using domain specific word embedding with deep learning and BERT," *IEEE Access*, vol. 9, pp. 106 363–106 374, 2021.

[5] S. Vujičić Stanković and M. Mladenović, "An approach to automatic classification of hate speech in sports domain on social media," *Journal of Big Data*, vol. 10, no. 1, pp. 1–16, 2023.

[6] S. Almohaimeed, S. Almohaimeed, A. A. Shafin, B. Carbunar, and L. Bölöni, "THOS: A benchmark dataset for targeted hate and offensive speech," in *Proc. of Data-centric Machine Learning Research (DMLR) Workshop at ICML 2023*, July 2023.

[7] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, "Predicting the type and target of offensive posts in social media," in *Proc. of NAACL-2019*, Jun. 2019, pp. 1415–1420.

[8] N. Ousidhoum, Z. Lin, H. Zhang, Y. Song, and D.-Y. Yeung, "Multilingual and multi-aspect hate speech analysis," in *Proc of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Nov. 2019, pp. 4675–4684.

[9] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A benchmark dataset for explainable hate speech detection," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 35, no. 17, 2021, pp. 14 867–14 875.

[10] N. Vashistha and A. Zubiaga, "Online multilingual hate speech detection: experimenting with Hindi and English social media," *Information*, vol. 12, no. 1, p. 5, 2020.

[11] E. W. Pamungkas and V. Patti, "Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon," in *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, Jul. 2019, pp. 363–370.

[12] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII*, H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Eds., 2020, pp. 928–940.

[13] V. Bonta, N. Kumaresh, and N. Janardhan, "A comprehensive study on lexicon based approaches for sentiment analysis," *Asian Journal of Computer Science and Technology*, vol. 8, no. S2, pp. 1–6, 2019.

[14] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proc. of the 5th Int. Workshop on Natural Language Processing for Social Media*, 2017, pp. 1–10.

[15] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, and D. Yang, "Latent hatred: A benchmark for understanding implicit hate speech," pp. 345–363, 2021.

[16] F. Huang, H. Kwak, and J. An, "Is ChatGPT better than human annotators? Potential and limitations of ChatGPT in explaining implicit hate speech," in *Companion Proc. of the ACM Web Conference 2023*, 2023, p. 294–297.

[17] 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," *arXiv preprint arXiv:1907.11692*, 2019.

[18] A. Gautam, P. Mathur, R. Gosangi, D. Mahata, R. Sawhney, and R. R. Shah, "#MeTooMA: Multi-aspect annotations of tweets related to the MeToo movement," in *Proc. of the Int. AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 209–216.