**Leveraging World Knowledge in Implicit Hate Speech Detection**

# Introduction

The main objective of the paper is to classify the implicit and explicit hate in the text. The proposed study mainly performed two tasks using the MLP as binary classification of text (hate speech or not hate speech), and secondly classification of hate speech into 6 subclasses. For the classification of the text, the proposed study extracted the embeddings from the pre-trained Bert model ‘bert-base-nli-mean-tokens’ and also extracted the knowledge description to concatenate with embeddings. The BERT model was trained with the combination of embeddings and knowledge description and only with model embeddings.

The proposed work attempt to detect implicit hate speech detection as it is harder than explicit hate speech detection in the text. But the subtype classification of implicit hate speech is still challenging. The subtype classification of implicit hate speech required more background knowledge for the proposer detection which is challenging for future studies.

# Related Work

The detection of hate speech has become a subject of great interest in recent times, and various approaches have been employed to address this issue. Earlier studies focused on identifying explicit abusive language through keyword-based techniques that relied on lexical features (Davidson et al., 2017; Waseem & Hovy, 2016). However, more recent research has emphasized the linguistic complexity and diversity of implicit hate expressions, which include stereotypes (Sap et al., 2020), indirect sarcasm, humor, and metaphor (Founta et al., 2018). These expressions cannot be identified by keyword-based systems alone. Implicit hate expressions can be just as damaging as explicit ones and contribute significantly to false negatives errors (Basile et al., 2019; Mozafari et al., 2020).

To detect implicit cases of hate speech, current solutions incorporate contextual information. For instance, (Gao & Huang, 2017) used original news articles as the context for the hateful comments. Other studies have created datasets with "implicit" labels or annotations (Caselli et al., 2020; ElSherief et al., 2021). This is essential not only for evaluation purposes but also for training, as systems that only rely on explicit features would fail to detect implicit hate speech, making them ineffective as moderation tools in real-world applications. In recent times, researchers have begun exploring the potential of incorporating real-world knowledge into related tasks, such as sarcasm detection, but not for hate speech detection. This line of research (Basu et al., 2021; Li et al., 2021) suggests that infusing commonsense knowledge into sarcasm detection models could improve the identification of sarcasm in cases where it is not obvious from the text. (Li et al., 2021) introduced a novel architecture that integrates knowledge into the learning model.

# Experiments

## Dataset

The proposed model utilized the Latent Hatred Dataset (ElSherief et al., 2021) in this study. The dataset comprises 21,480 tweets from the most notable extremist groups in the United States, with 7,100 of these tweets classified as implicit hate speech and 1,089 as explicit hate speech. The implicit hate tweets were grouped into six categories based on the classification system presented in Table 1.

|  |  |
| --- | --- |
| Class | Samples |
| Grievance | 1538 |
| Incitement | 1269 |
| Inferiority | 863 |
| Irony | 797 |
| Stereotypes | 1133 |
| Threats | 666 |

The statistics of the used dataset are not similar to the statistics reported in the paper. It is probably due to the upgraded version of the dataset. Collectively the structure and the classes of the dataset are still similar.

## Implementation

The implementation of the paper for MLP classification was performed from starch without using the GitHub repo. Although, the paper also uses the Radboud entity Linker, but it is not available right now.

## Results

The paper implemented the MLP classifier firstly for binary classification with the combination of model embeddings and background knowledge or with model embeddings only. For the model embeddings only, model showed the 0.7247% accuracy while the model showed 0.7232% accuracy with the combination of model embeddings and background knowledge. The confusion matrix of both models is shown below:

|  |  |
| --- | --- |
|  |  |

The MLP model was also trained with a combination of model embeddings and background knowledge or model embeddings only for the subtype classification of implicit hate speech. The confusion matrix of both models is also shown below figures. The below table also showed the results of all models compared to the results reported in the published paper.

|  |  |
| --- | --- |
|  |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Reproduced | | | | Reported | | | |
| Classification | Feature Vector | Accuracy | Precision | Recall | F1 Score | Accuracy | Precision | Recall | F1 Score |
| Binary | Embeddings | 0.7274 | 0.71 | 0.70 | 0.70 | 65 % ± 1.5 % | 65 % ± 1.5 % | 65 % ± 1.5 % | 65 % ± 1.5 % |
| Binary | Embeddings + BK | 0.7232 | 0.71 | 0.69 | 0.69 | 75 % ± 1.4 % | 75 % ± 1.4 % | 75 % ± 1.4 % | 75 % ± 1.4 % |
| Multiclass | Embeddings | 0.5334 | 0.53 | 0.54 | 0.53 | 52 % ± 1.3 % | 52 % ± 1.3 % | 52 % ± 1.3 % | 54 % ± 1.3 % |
| Multiclass | Embeddings + BK | 0.5127 | 0.50 | 0.51 | 0.50 | 42 % ± 1.3 % | 40 % ± 1.3 % | 41 % ± 1.3 % | 44 % ± 1.3 % |

## Discussion

Firstly, the Entity linker model used by the published paper is no longer available on GitHub or via API. However, the reproduced results for binary classification and multiclass classification are better than the reported results. The main factor of these is probably the upgradation of the dataset. As the dataset has fewer samples at the time of paper publication but now the dataset has more samples that can contribute to the betterment of the results. Moreover, very few hyperparameters of MLP are discussed in the published paper while we tuned many hyperparameters that can be a factor in the significant results.

As the binary and multiclass models perform well compared to the published models only for the tweet text embedding. We can hypothesized that these models will perform well with the combination of tweets embeddings and background knowledge if the Linker model code or API is available in future.

## Error Analysis

For the error analysis, the confusion matrix of the model on the test set was plotted that showed the total accurate samples of each class and the false negative and positive samples of each class in the above figures. Moreover, the author of the paper perform post processing after linker model in which some words that were associated with hate speech labeled as neutral that increase the results of classification.

# Conclusion

The report presents the results reproduced using the methodology and experiments of the selected paper. Currently, the results of the selected paper are partially reproducible. If the linker model is available in future, then all results of the paper can be reproduced easily. However, the reproduced results validate that the reported results are true and well-defined.

# References

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *NAACL HLT 2019 - International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, 54–63. https://doi.org/10.18653/V1/S19-2007

Basu, S., Chowdhury, R., & Chaturvedi, S. (2021). *Does Commonsense help in detecting Sarcasm?* 9–15. https://doi.org/10.18653/v1/2021.insights-1.2

Caselli, T., Basile, V., Mitrovi´cmitrovi´c, J., Kartoziya, I., & Granitzer, M. (2020). *I Feel Offended, Don’t Be Abusive! Implicit/Explicit Messages in Offensive and Abusive Language* (pp. 6193–6202). https://aclanthology.org/2020.lrec-1.760

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*, 512–515. https://doi.org/10.1609/icwsm.v11i1.14955

ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., & Yang, D. (2021). Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 345–363. https://doi.org/10.18653/V1/2021.EMNLP-MAIN.29

Founta, A. M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, 491–500. https://doi.org/10.1609/icwsm.v12i1.14991

Gao, L., & Huang, R. (2017). *Detecting Online Hate Speech Using Context Aware Models*. 260–266. https://doi.org/10.26615/978-954-452-049-6\_036

Li, J., Pan, H., Lin, Z., Fu, P., & Wang, W. (2021). Sarcasm Detection with Commonsense Knowledge. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *29*, 3192–3201. https://doi.org/10.1109/TASLP.2021.3120601

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, *15*(8), e0237861. https://doi.org/10.1371/JOURNAL.PONE.0237861

Sap, M., Gabriel, S., Qin, L., Smith, D. J. N. A., & Choi, Y. (2020). Social Bias Frames: Reasoning about Social and Power Implications of Language. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 5477–5490. https://doi.org/10.18653/V1/2020.ACL-MAIN.486

Waseem, Z., & Hovy, D. (2016). Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *HLT-NAACL 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Student Research Workshop*, 88–93. https://doi.org/10.18653/V1/N16-2013