

HATE SPEECH DETECTION ON PLATFORM X USING WORD2VEC AND CONVOLUTIONAL NEURAL NETWORK (CNN)

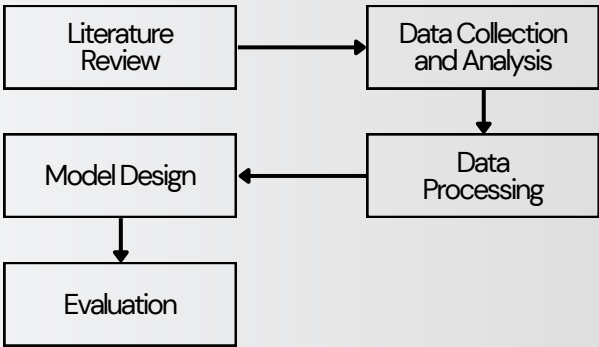
Adhy Ardana Setyawan, Aditiyawan

BACKGROUND



- Hate speech, particularly in digital spaces like social media, is rapidly increasing and often targets vulnerable groups, posing serious threats to social harmony, youth safety, and democratic values. In Indonesia, this phenomenon has escalated significantly, while existing moderation efforts remain insufficient.
- To address this challenge, a machine learning approach using a Convolutional Neural Network (CNN) combined with Word2Vec embeddings is proposed. This method offers an effective and scalable solution for detecting hate speech in Indonesian by capturing complex language patterns and contextual nuances.

RESEARCH METHODOLOGY



RESULT AND DISCUSSION

EVALUATION METRICS ACROSS SPLIT SCENARIOS

Split	Class	Precision	Recall	F1-Score	Accuracy
80:20	HS	0.88	0.86	0.87	0.88
	Non HS	0.89	0.91	0.90	
70:30	HS	0.90	0.80	0.85	0.87
	Non HS	0.85	0.92	0.89	
60:40	HS	0.88	0.78	0.83	0.85
	Non HS	0.84	0.92	0.88	

COMPARISON WITH VS. WITHOUT WORD2VEC

Method	Class	Precision	Recall	F1-Score	Accuracy
With Word2Vec	HS	0.88	0.86	0.87	0.88
	Non HS	0.89	0.91	0.90	
Without Word2Vec	HS	0.86	0.68	0.76	0.81
	Non HS	0.78	0.91	0.84	

CONCLUSION



- This research successfully developed a hate speech detection model for the Indonesian language using a Convolutional Neural Network (CNN) combined with Word2Vec embedding.
- The model achieved an overall accuracy of 88.44%, with balanced performance across both classes. For non-hate speech, it reached 88% precision, 86% recall, and an F1-score of 87%. For hate speech, the model performed slightly better, with 89% precision, 91% recall, and a 90% F1-score.

REFERENCE



- [1] S. Riyadi, A. D. Andriyani, A. M. Masyhur, C. Damarjati, and M. I. Solihin, "Detection of Indonesian hate speech on Twitter using hybrid CNN-RNN," in Proc. 2023 Int. Conf. on Information Technology and Computing (ICITCOM), IEEE, Dec. 2023, pp. 352–356.
- [2] N. M. Andini, Y. Findawati, I. R. I. Astutik, and A. Eviyanti, "Implementasi convolutional neural network (CNN) untuk mendeteksi ujaran kebencian dan emosi di Twitter," SMATIKA JURNAL, vol. 14, pp. 314–325, Dec. 2024. [Online]. Available: <https://jurnal.stiki.ac.id/SMATIKA/article/view/1346>
- [3] N. S. Mullah and W. M. N. W. Zainon, "Advances in machine learning algorithms for hate speech detection in social media: A review," pp. 88364–88376, 2021.