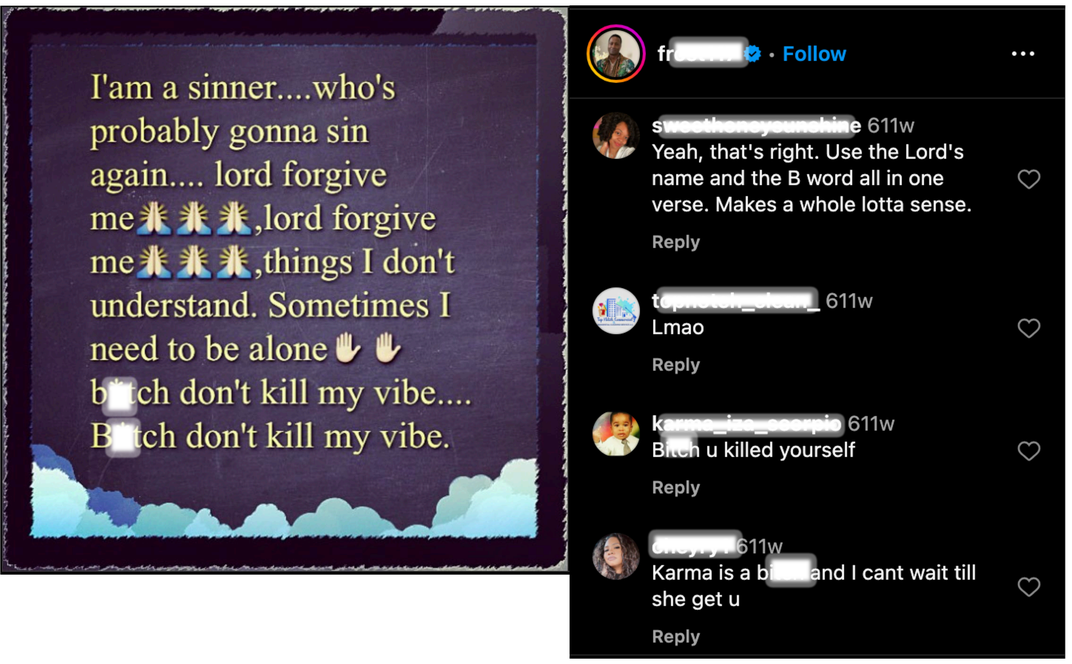
Visual BullyBlocker: Extending Cyberbullying Detection to Images

Students: Rehonoma Hasan Jahin, Alisa Samakhval, Manuel Sandoval Madrigal
Faculty members: Daniel Moreira, Yasin Silva

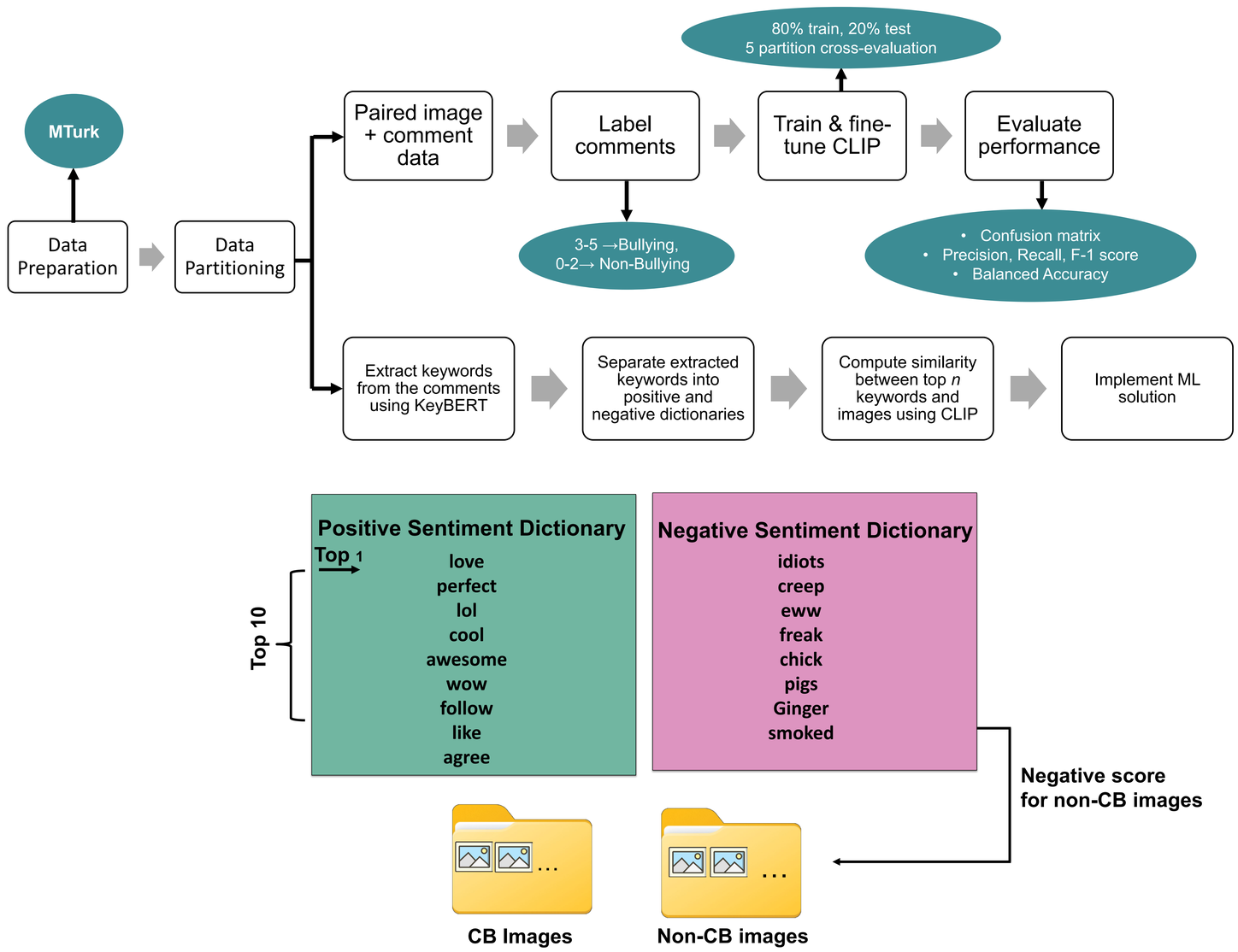
Introduction



Background & Objectives

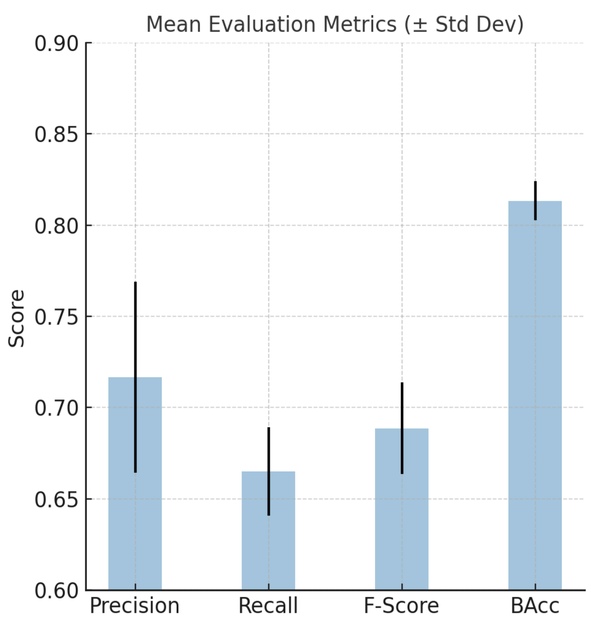
- BullyBlocker is a prior development that detects cyberbullying by analyzing social media text.
- Use CLIP to analyze paired Instagram comments and images, identifying bullying with both visual and textual signals.
- Develop and refine a sentiment analysis-driven approach for detecting cyberbullying in images.

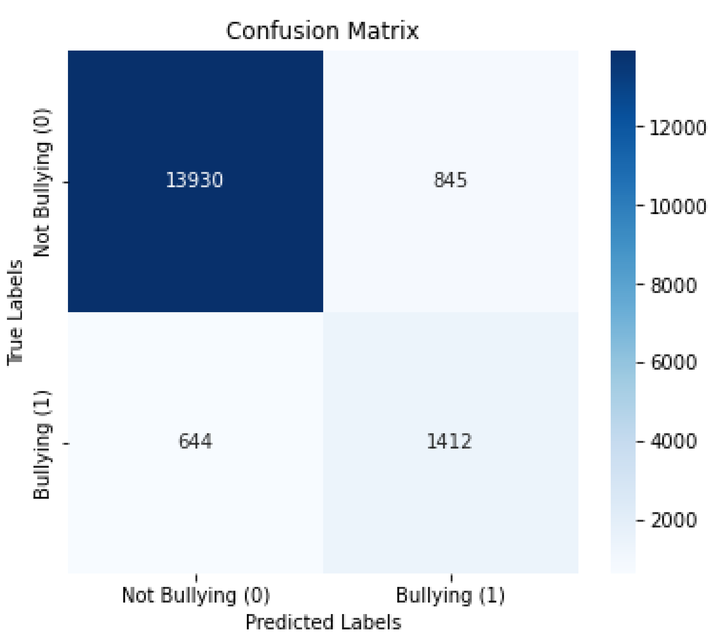
Methods



Results

Method 1:





Method 2:

CB Images		Non-CB Images	
Positive Score	Negative Score	Positive Score	Negative score
22.75	9.59	28.74	4.11

Future Work

- Perform multimodal classification instead of binary classification
- Incorporate insights from psychological research on language patterns

References

- Y. Silva, D. Hall, C. Rich. BullyBlocker: Toward an Interdisciplinary Approach to Identify Cyberbullying. Social Network Analysis and Mining (SNAM), 8, 1, 2018. DOI: 10.1007/s13278-018-0496-z .
- A. Radford, J.-W. Kim, C. Hallacy, A. Ramesh, G. Goh, S, Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark., et al. Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning (ICML), 2021.
- M. Grootendorst, Keybert: Minimal keyword extraction with BERT, 2020, http://dx.doi.org/10.5281/zenodo.4461265.

BullyBlocker

Acknowledgments
This work was supported by a LUC Provost Fellowship (Rehonoma Hasan Jahin), and NSF Award #2227488.
Thank you!