**Project Title:**

**Identifying Hate Speech and Offensive Language On Twitter**

**Group members:**

Erica Zhou [xuanz1@umbc.edu](mailto:xuanz1@umbc.edu)

Neil Kpamegan [kneil1@umbc.edu](mailto:kneil1@umbc.edu)

John Galpin [jgalpin1@umbc.edu](mailto:jgalpin1@umbc.edu)

Reenah Sheikh [rsheikh1@umbc.edu](mailto:rsheikh1@umbc.edu)

**Project Objectives:**

Our goal is to determine and implement the best algorithm for the classification of hate speech and offensive language on Twitter. Using a labeled dataset we plan to use text similarity metrics and classification techniques. The dataset will be split into training and test sets. Secondary testing will be done on unlabeled archived sources. The second half of the project will be manually labeling the unlabeled data. From there we can determine the reliability of similarity metrics to classify unlabeled data.

**Data Set Description:**

The labeled data set contains over 24000 tweets. Each labeled Hate Speech, Offensive or Neither. There are columns that count the number of votes by humans for each classification. The second data set is an unlabeled archived twitter stream dataset.

**Datasets**

<https://data.world/thomasrdavidson/hate-speech-and-offensive-language>

<https://archive.org/search.php?query=collection%3Atwitterstream&sort=-publicdate>\

**Key LIterature:**

Papers

<https://ieeexplore.ieee.org/abstract/document/8292838>

Watanabe, H., Bouazizi, M., & Ohtsuki, T. (2018). Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE Access*, *6*, 13825-13835.

<https://dl.acm.org/citation.cfm?id=3054223>

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion* (pp. 759-760). International World Wide Web Conferences Steering Committee.

<https://link.springer.com/chapter/10.1007/978-3-319-93417-4_48>

Zhang, Z., Robinson, D., & Tepper, J. (2018, June). Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference* (pp. 745-760). Springer, Cham.

<https://dl.acm.org/citation.cfm?id=3054223>

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., & Bhamidipati, N. (2015, May). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web* (pp. 29-30). ACM.

<https://dl.acm.org/citation.cfm?id=3078723>

Mondal, M., Silva, L. A., & Benevenuto, F. (2017, July). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media* (pp. 85-94). ACM.

<https://dl.acm.org/citation.cfm?id=3270103>

Gröndahl, T., Pajola, L., Juuti, M., Conti, M., & Asokan, N. (2018, October). All You Need is: Evading Hate Speech Detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security* (pp. 2-12). ACM.

Possible Code Base

<https://github.com/eliorc/Medium/blob/master/MaLSTM.ipynb>

<https://ai.stackexchange.com/questions/6297/can-we-apply-neural-networks-to-two-text-comparisons>

Siamese Manhattan LSTM Implementation

<https://medium.com/mlreview/implementing-malstm-on-kaggles-quora-question-pairs-competition-8b31b0b16a07>

Rating Text Similarity Methodologies

<https://medium.com/@adriensieg/text-similarities-da019229c894>