

Detecting Cyberbullying and Hate Speech on Twitter

Pranav Dahiya

Shiv Nadar University

August 11, 2020

- 1 Introduction
 - Problem
 - Aim
 - Data
- 2 Preprocessing
- 3 Feature Extraction
- 4 Results
 - Baseline
 - Cross Validation
- 5 Discussion
- 6 Future Work

Problem

- In September 2011, Jamey Rodemayer, an American boy of fourteen, hanged himself after being subjected to cyber-bullying for years because of his sexual orientation. (Wingate, Minney, and Guadagno)
- In October 2012, the young Canadian Amanda Todd, committed suicide due to constant bullying, physically and online. (Lester, McSwain, and Gunn III)
- In January 2018, the young Australian Teenager Dolly Everett committed suicide after becoming a victim of cyberbullying. (Kennedy and Coulter)

Aim

Build a robust hate speech and cyberbullying classifier for Twitter

Why Twitter?

- More than 350 million active users as of 2018 (Mody et al.)
- Very popular among adolescents

Data

- Dataset for Detection of CyberTrolls (DataTurks)
 - 20000 tweets, 2 classes
 - 1: Instances of Cyberbullying (39%)
 - 0: Normal tweets (71%)
- Twitter Hatespeech Dataset (Davidson et al.)
 - 25000 tweets, 3 classes
 - 0: Hate Speech (5.7%)
 - 1: Offensive tweets but not hate speech (76.7%)
 - 2: Normal tweets (17.6%)

Preprocessing

- 1 Remove urls and replace with the string "url"

Preprocessing

- 1 Remove urls and replace with the string "url"
- 2 Decode unicode characters and html tags

Preprocessing

- 1 Remove urls and replace with the string "url"
- 2 Decode unicode characters and html tags
- 3 Replace emojis and emoticons with strings corresponding to their meaning

Preprocessing

- 1 Remove urls and replace with the string "url"
- 2 Decode unicode characters and html tags
- 3 Replace emojis and emoticons with strings corresponding to their meaning
- 4 Create TfIdf vectors from bigrams and trigrams.

Feature Extraction

- Profanity features (Zhou)
- Emojis (NeelShah18)
- Character based features
- Number of pronouns
- Length
- Sentiment (Salawu, He, and Lumsden Di Capua, Di Nardo, and Petrosino)
- Regression trained on tfidf vectors

Baseline

Table: Baseline Confusion Matrix for Dataset 1

	0	1
0	0.996	0.004
1	0.048	0.951

Table: Baseline Results for Dataset 1

Accuracy	0.97 ± 0.01
Precision	0.98 ± 0.01
F1	0.98 ± 0.01
Recall	0.98 ± 0.01

Baseline

Table: Baseline Confusion Matrix for Dataset 2

	0	1	2
0	0.06	0.84	0.10
1	0.00	0.97	0.03
2	0.00	0.15	0.85

Table: Baseline Results for Dataset 2

Accuracy	0.62 ± 0.01
Precision	0.89 ± 0.02
F1	0.88 ± 0.01
Recall	0.90 ± 0.01

Results

Table: Confusion Matrix

	0	1	2
0	0.978	0.021	0.001
1	0.001	0.998	0.001
2	0.000	0.006	0.993

Table: Results of SVM on twitter dataset

Accuracy	0.99 ± 0.01
Precision	1.00 ± 0.00
F1	1.00 ± 0.00
Recall	1.00 ± 0.00

Discussion

- Tried combination of feature sets found in literature

Discussion

- Tried combination of feature sets found in literature
- Novelty in robustness of pre-trained models used for generating features

Discussion

- Tried combination of feature sets found in literature
- Novelty in robustness of pre-trained models used for generating features
- Novelty in preprocessing with emojis and bigram/trigram regression

Challenges

- Lack of header and user information in available datasets.
- Lack of large quality dataset to attempt semi-supervised learning.
- Use of semantic features as by Verma and Hossain for phishing email detection might be useful here. However, no improvement headroom left in available datasets to evaluate it.



DataTurks. “Tweets Dataset for Detection of Cyber-Trolls”. 2018. Web.



Di Capua, M., E. Di Nardo, and A. Petrosino. “Unsupervised cyber bullying detection in social networks”. 2016 *23rd International Conference on Pattern Recognition (ICPR)*. Dec. 2016. 432–437. Print.



Lester, David, Stephanie McSwain, and John Gunn III. “Suicide and the Internet: the case of Amanda Todd”. *International journal of emergency mental health* 15 (Jan. 2013): 179–80. Print.



Mody, A., et al. “Identification of Potential Cyber Bullying Tweets using Hybrid Approach in Sentiment Analysis”. 2018 *International Conference on Electrical, Electronics,*

Communication, Computer, and Optimization Techniques (ICEECOT). Dec. 2018. 878–881. Print.



NeelShah18. "Emot". 2020. Web.

Salawu, S., Y. He, and J. Lumsden. "Approaches to Automated Detection of Cyberbullying: A Survey". *IEEE Transactions on Affective Computing* 11.1 (Jan. 2020): 3–24. Print.



Verma, Rakesh and Nabil Hossain. "Semantic Feature Selection for Text with Application to Phishing Email Detection". *Information Security and Cryptology – ICISC 2013*. Edited by Hyang-Sook Lee and Dong-Guk Han. Cham: Springer International Publishing, 2014. 455–468. Print.



Wingate, Victoria, Jessica Minney, and Rosanna Guadagno.
 “Sticks and stones may break your bones, but words will always hurt you: A review of Cyberbullying”. *Social Influence* 8 (Apr. 2013). Print.



Zhou, Victor. "Building a Better Profanity Detection Library with scikit-learn". 2019. Web.