Analyzing Adversarial Strategies and Countermeasures for Cyberbullying Detection

Students: Maddie Juarez, Manny Sandoval Madrigal, Eldor Abdukhamidov

Advisors: Mohammed Abuhamad, Yasin N. Silva, Tamer Abuhmed

Abstract

- Cyberbullying has harmful effects on users and is increasing on social networking sites, with messages spreading rapidly
- Cyberbullying systems are vulnerable to adversarial attacks, such as TextFooler
- We evaluated the robustness of cyberbullying detection models on traditional machine learning (ML) and newer LLM approaches
- Used real-world datasets from Instagram, Twitter, and Vine
- Adversarial attacks are found to significantly reduce the accuracy of detection models
- Future work would expand on adversarial training to improve robustness

Background & Framework

- **Cyberbullying Detection Models**
  - Support Vector Machines (SVM)
  - Random Forest (RF)
  - K-Nearest Neighbors (KNN)
  - Naïve Bayes (NB)
  - XGBoost
  - Long Short Term Memory (LSTM)
  - Bi-directional Long Short Term Memory (Bi-LSTM)
  - Bi-directional Gated Recurrent Unit (Bi-GRU)
  - Gated Recurrent Unit (GRU)
  - Bidirectional Encoder Representations from Transformers (BERT)
  - a fine-tuned version of BERT for cyberbullying detection (CyberBERT)

Data Representation Methods

- Term Frequency-Inverse Document Frequency (**TF-IDF**)
  - Character-level
  - Word-level *n*-grams (1-3 *n*-grams)
- Pre-trained word embeddings like **Word2Vec**
- Pretrained language models (**BERT**)

Baseline Performance of Cyberbullying Detection Models:

Method	Features	Instagram			Twitter			Vine		
		F1	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**
XGBoost	TF-IDF 1000	0.7282	0.8257	0.6528	**0.7638**	0.7891	<u>0.7403</u>	0.6047	0.7355	0.5462
	TF-IDF *bi*-gram	0.6710	0.8238	0.5666	0.2630	0.7730	0.1596	0.5592	0.6333	0.5069
	TF-IDF Char	0.6692	0.7825	0.5864	0.7259	0.8021	0.6632	0.5889	0.6907	0.5069
	Count 1000	<u>0.7290</u>	0.8027	0.6702	0.7639	0.7942	0.7361	0.6050	0.7012	0.5396
RF	TF-IDF *bi*-gram	0.6506	<u>0.8797</u>	0.5543	0.3397	0.2449	0.5546	0.5856	0.7392	0.4870
	TF-IDF Char	0.5055	0.8678	0.3571	0.3353	0.2183	0.7239	0.4584	<u>0.7995</u>	0.3294
	TF-IDF 1000	0.6524	**0.9016**	0.5123	0.4376	0.2965	0.6950	0.5275	0.7271	0.4464
	Count 1000	0.5743	0.8098	0.4457	0.4463	0.3345	0.6701	0.5744	0.6735	0.5035
SVM	TF-IDF *bi*-gram	0.6848	0.8500	0.5916	0.2249	0.7729	0.1316	0.5021	**0.8046**	0.3688
	TF-IDF Char	0.5828	0.7861	0.4631	0.4334	0.3077	0.7325	0.6256	0.7224	0.5527
	TF-IDF 1000	0.7139	0.8318	0.6283	<u>0.7520</u>	<u>0.8238</u>	0.6917	0.6177	0.7583	0.5157
	Count 1000	0.5898	0.8209	0.4606	0.4239	0.2999	0.7233	0.5473	0.7654	0.4278
NB	TF-IDF *bi*-gram	0.5069	0.3762	<u>0.7787</u>	0.5498	**0.8793**	0.3999	0.5044	0.3717	**0.8971**
	TF-IDF Char	0.3640	0.2388	0.7662	0.4397	0.3731	0.5831	0.5454	0.4342	0.7336
	TF-IDF 1000	0.6180	0.4256	**0.8082**	0.4775	0.4442	0.5164	0.6021	0.4792	0.7208
	Count 1000	0.4309	0.4141	0.4506	0.4651	0.4427	0.4901	0.4197	0.5168	0.3555
KNN	TF-IDF *bi*-gram	0.4458	0.5123	0.3947	0.3716	0.5312	0.2857	0.5616	0.4824	0.6721
	TF-IDF Char	0.6301	0.6053	0.6571	0.3764	0.5661	0.2819	<u>0.6538</u>	0.5368	<u>0.8361</u>
	TF-IDF 1000	0.5588	0.5758	0.5429	0.2130	0.4386	0.1406	0.5930	0.4595	<u>0.8361</u>
	Count 1000	0.4314	0.6875	0.3143	0.3384	0.4676	0.2652	0.5743	0.7250	0.4754
LSTM	Embedding Word2Vec	0.3421	0.4530	0.2813	0.4963	0.4893	0.5035	0.5097	0.4780	0.5655
GRU	Embedding (Trainable)	0.6848	0.7421	0.6358	0.5034	0.5240	0.4852	0.6525	0.6821	0.6254
CyberBERT	Transformer Embeddings	**0.7477**	0.7612	0.7347	0.7493	0.7134	**0.7892**	**0.7387**	0.7523	0.7256

Results

Character-level Attacks

- Homoglyph substitutions (e.g., "S" -> "$")
- Evades traditional token-based defenses

1. **Greedy word swaps with weighted saliency scores**
   - Measures individual contribution to the model's prediction and its interactions with surrounding words
2. **Greedy homoglyph substitution**
   - Generate similar glyphs for words with high saliency
   - Test each substitution by evaluating model predictions
   - Accept changes that flip prediction or maximally reduce confidence
   - Apply greedy selection for optimal character replacements
3. **Iterate through Important Words**
   - Process words in order of saliency score
   - Preserves changes that reduce model confidence
   - Continues until prediction flips or all key words are tested

	Example	Prediction
B	*—So precious is smart or nah rem-miem etaylorrose—*	**Non-Bullying** (76%)
A	*—So precious is smart or naη rem-miem etaylorrose—*	**Bullying** (74%)
B	*—You can only see the symbolism blatantly shown in your face—*	**Non-Bullying** (73%)
A	*—Yo∪ can only see the symbolism blatantly shown in your face—*	**Bullying** (66%)
B	*—you dont got a profile pic cuz u ugly as f[redacted]k—*	**Bullying** (87%)
A	*—you dont got a profile pic cuz u ugly as f[redacted]κ—*	**Non-Bullying** (68%)

Models	Features	Instagram			Twitter			Vine		
		Success Rate	# Query	Avg. chars	Success Rate	# Query	Avg. chars	Success Rate	# Query	Avg. chars
XGBoost	TF-IDF 1000	0.07	5183	3.18	0.19	62	3.99	0.16	1160	3.25
	TF-IDF Char	<u>0.09</u>	4655	7.45	<u>0.35</u>	113	5.72	0.28	1236	5.51
	Count 1000	0.06	4421	2.83	0.16	78	4.63	0.21	1508	3.37
	TF-IDF *bi*-gram	0.07	5948	3.83	0.07	73	4.11	0.10	1490	1.70
SVM	TF-IDF 1000	0.04	3549	43.09	0.18	66	5.23	0.30	2401	25.64
	TF-IDF Char	0.13	1643	47.79	0.28	83	8.29	0.24	1384	38.30
	Count 1000	0.02	4382	76.83	0.13	53	4.78	0.10	1325	15.42
	TF-IDF *bi*-gram	0.04	3645	44.06	0.09	74	4.84	0.12	1630	70.12
Random Forest	TF-IDF 1000	0.01	3266	6.25	0.19	38	5.28	0.08	1413	7.73
	TF-IDF Char	0.00	-	-	0.13	65	5.68	0.05	877	7.00
	Count 1000	0.01	3527	8.45	0.17	57	4.27	0.08	1689	4.00
	TF-IDF *bi*-gram	0.03	5944	8.45	0.17	57	4.27	0.08	1689	4.00
Naive Bayes	TF-IDF 1000	**0.44**	980	19.14	0.32	90	5.80	0.16	894	30.61
	TF-IDF Char	0.01	2077	77.65	0.31	53	6.82	**0.41**	616	95.21
	Count 1000	0.07	4460	22.43	0.13	92	5.72	0.05	2490	22.80
	TF-IDF *bi*-gram	0.41	2357	2.00	0.13	68	6.42	<u>0.39</u>	497	39.26
KNN	TF-IDF 1000	0.003	3625	1.00	0.32	84	5.63	0.08	561	1.53
	TF-IDF Char	0.02	3831	1.20	0.32	82	5.16	0.03	481	1.00
	Count 1000	0.003	3278	1.00	0.17	42	4.66	0.03	1102	1.00
	TF-IDF *bi*-gram	0.02	2267	1.14	**0.49**	94	5.01	0.12	572	1.48
LSTM	Embeddings	–	–	–	–	–	–	–	–	–
GRU	Embeddings (Trainable)	0.08	766	4.64	0.42	77	5.89	<u>0.39</u>	1418	5.59
CyberBERT	Transformer Embeddings	–	–	–	0.10	126	8.20	0.08	1400	43.75

Word-level Attacks

- Aims to generate adversarial examples that change the model's predictions while maintaining the original context and meaning

1. **Compute Importance Scores**
   - Measured by how the model's prediction probability changes when each word is removed or masked
2. **Iterative Substitution**
   - Finds semantically similar synonyms using pre-trained word embeddings (e.g., Word2Vec GloVe) for each word starting with the most important one
3. **Proceed to Next Important Word**
   - Algorithm maintains substitution that reduce prediction probability until either the model's prediction changes or all important words have been processed

Original Text: *"Every single NFL player should be kneeling this Sunday Every Single One Dont let this POS President get away wthis sh."*
Model Prediction: Negative
Importance Ranking: Low (–) (+) High

Substitutions:
"POS" → "awful" → **Prediction:** Negative (lower probability)
"POS" → "controversial" → **Prediction:** <u>Positive</u> (prediction flip)

Adversarial Example: *"Every single NFL player should be kneeling this Sunday Every Single One Dont let this controversial President get away wthis sh."*
New Prediction: <u>Positive</u>

Models	Features	Instagram			Twitter			Vine		
		Success Rate	# Query	Avg. words	Success Rate	# Query	Avg. words	Success Rate	# Query	Avg. words
XGBoost	TF-IDF 1000	0.41	4561	4.35	0.38	155	2.00	0.57	987	4.00
	TF-IDF Char	0.37	1291	7.48	0.61	53	4.15	0.59	334	7.32
	Count 1000	0.45	4160	6.83	0.39	90	3.16	0.60	785	5.68
	TF-IDF *bi*-gram	0.19	4436	8.31	0.11	117	5.52	0.33	768	8.23
SVM	TF-IDF 1000	0.19	10226	30.12	0.30	43	7.00	0.65	1675	56.75
	TF-IDF Char	<u>0.82</u>	1471	28.44	<u>0.62</u>	90	10.00	0.63	538	23.31
	Count 1000	0.04	1994	21.00	0.23	70	4.00	0.29	1594	96.67
	TF-IDF *bi*-gram	0.06	8178	39.57	0.13	57	5.50	0.25	7178	157.48
Random Forest	TF-IDF 1000	0.03	2030	8.12	0.28	163	2.00	0.15	860	41.54
	TF-IDF Char	0.01	8272	6.70	0.34	123	3.00	0.18	794	24.36
	Count 1000	0.02	2493	4.94	0.25	170	5.00	0.12	762	10.00
	TF-IDF *bi*-gram	0.04	1243	10.50	0.29	55	7.00	0.09	570	15.00
Naive Bayes	TF-IDF 1000	0.49	276	5.50	0.49	160	4.50	0.35	838	62.00
	TF-IDF Char	0.45	589	7.50	0.59	72	12.00	0.48	458	10.00
	Count 1000	0.18	1928	10.20	0.29	130	4.00	0.11	855	86.35
	TF-IDF *bi*-gram	0.01	2015	11.40	0.19	61	2.00	0.49	246	28.00
KNN	TF-IDF 1000	0.03	1670	3.40	0.14	72	3.50	0.10	299	3.37
	TF-IDF Char	0.04	733	2.92	0.36	69	3.58	0.10	224	2.50
	Count 1000	0.01	629	2.00	0.30	72	2.98	0.06	674	2.18
	TF-IDF *bi*-gram	0.06	1341	3.53	0.16	78	4.12	0.25	533	3.31
LSTM	Embedding Word2Vec	0.73	545	10.80	0.59	84	2.37	0.64	273	9.00
GRU	Embeddings (Trainable)	**0.88**	2568	14.26	**0.87**	55	6.00	**0.88**	985	18.05
CyberBERT	Transformer Embeddings	0.12	1750	7.75	0.15	142	6.20	0.10	990	19.00

BullyBlocker

This work was supported by National Science Foundation Award # 2227488