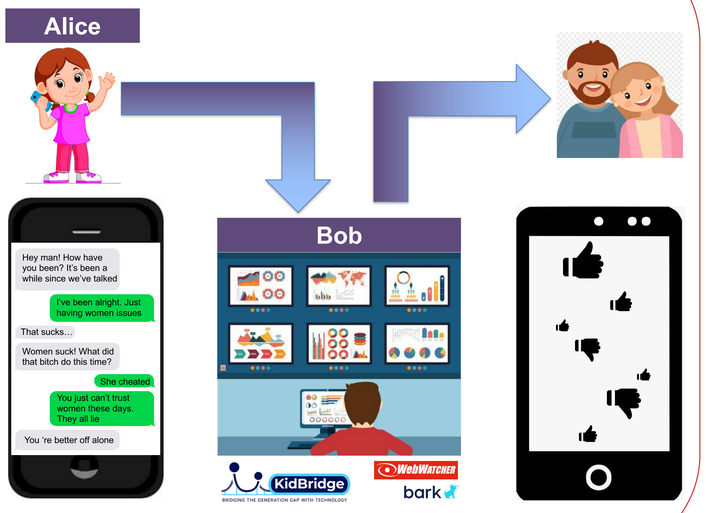
Privacy-Preserving Classification of Personal Text Messages with Secure Multi-Party Computation

Devin Reich[1], Ariel Todoki[1], Rafael Dowsley[2], Anderson Nascimento[1]and Martine De Cock[1]
reichd@uw.edu, atodoki@uw.edu, rafael@dowsley.net, andclay@uw.edu, mdecock@uw.edu
[1]School of Engineering and Technology, University of Washington Tacoma
[2]Department of Computer Science, Bar-Ilan University

Problem : Oblivious Text Classification

Motivating Example: Parental Control



A lot of personal information is leaked

Oblivious Text Classification

Alice wishes to classify a personal text with:

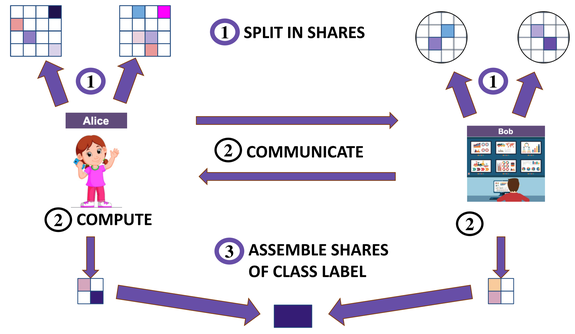
- a pre-trained text classifier held by Bob
- while keeping Alice's personal text hidden from Bob and
- keeping Bob's classifier hidden from Alice

Research Objective

- Develop efficient and provably secure protocol for oblivious classification of unstructured text.
- Almost all existing work is for secure classification of *structured* information extracted from texts, assuming that Alice extracts features on her end. This leaks information to Alice on what features Bob needs. Our approach privately extracts from Alice's text the unigrams and bigrams that Bob needs, without Bob having to show these to Alice.
- Only other existing approach on secure classification of unstructured text uses homomorphic encryption (HE) and takes 19 min to classify a tweet. [1]

Secure Protocol

Experimental setup



- Honest but Curious
- All computations done over additive shares
- No information is leaked about what the actual data is

Privacy Preserving Text Classification

Protocolπ_{TC} :

- Alice and Bob execute the secure feature extraction protocolπ_{FE} with input sets A and B in order to obtain the secret shares $[X]_2$ of the feature vector
- They run the protocol π_{2toQ} (secure conversion between Z_2 and Z_q) to obtain shares $[X_i]_q$

Protocolπ_{TC} - LR:

- Alice and Bob run the secure logistic regression classification protocol π_{LR} in order to get the result of the classification. The LR model M is given as input to π_{LR} by Bob, and the secret shared feature vector x by both of them. Bob gets the result of the classification M(x).

Protocolπ_{TC} - AB:

- Alice and Bob run the secure AdaBoost classification protocol π_{AB} to obtain the result of the classification. The secret shared feature vector x is given as input to π_{AB} by both of them, and the two weighted probability vectors $y = (y_{1,0}, y_{1,1}, \dots y_{n,0}, y_{n,1})$ and $z = (z_{1,0}, z_{1,1}, \dots z_{n,0}, z_{n,1})$ that constitute the model are specified by Bob. Bob gets the output class c.

Results

Experimental setup

- Binary classification task: label a message as hateful or not
- Data from HatEval competition at SemEval 2019.
- 10,000 tweets; 60% annotated as hate speech
- We converted all characters to lowercase, and turned each tweet into a set of word unigrams and bigrams.
- Trained Adaboost and Logistic Regression models have an accuracy at par with those of winner of competition [2].
- Experiments ran on AWS c5.9xlarge machines with 36 vCPUs, 72.0 GiB Memory. Each of the parties ran on separate machines (connected with a Gigabit Ethernet network). Results cover communication time in addition to computation time.

	Unigrams				Unigrams+Bigrams			
	Acc	Time (in sec)			Acc	Time (in sec)		
		Extr	Class	Tot		Extr	Class	Tot
Ada; 50 trees; depth 1	71.6%	0.8	6.4	7.2	73.3%	1.5	6.6	8.1
Ada; 200 trees; depth 1	73.0%	2.8	6.4	9.2	74.2%	9.4	6.6	16.0
Ada; 500 trees; depth 1	73.9%	6.6	6.7	13.3	**74.4%**	**21.6**	**6.7**	**28.3**
Logistic regression (50 feat.)	**72.4%**	**0.8**	**3.7**	**4.5**	73.8%	1.5	3.8	5.3
Logistic regression (200 feat.)	73.3%	2.8	3.7	6.5	73.7%	9.4	3.8	13.2
Logistic regression (500 feat.)	73.4%	6.6	3.8	10.4	74.2%	21.6	4.1	25.7
Logistic regression (all feat.)	73.1%	318.0	6.1	324.1	73.8%	5,371.9	24.9	5,396.8

Accuracy (Acc) results using 5-fold cross-validation over the corpus of 10,000 tweets. Total time (Tot) needed to securely classify a text with our framework, broken down in time needed for feature vector extraction (Extr) and time for feature vector classification (Class).

Conclusion

- First efficient provably secure method for private classification of unstructured text
- No loss in accuracy compared to in the clear
- Runs in under 5 sec while previous method took 19 min [1]

References

[1] G. Costantino, A. La Marra, F. Martinelli, A. Saracino, and M. Sheikhalishahi. Privacy-preserving text mining as a service. In 2017 IEEE Symposium on Computers and Communications (ISCC), p. 890–897, 2017.
[2] V. Indurthi, B. Syed, M. Shrivastava, N. Chakravartula, M. Gupta, M., V. Varma. FERMI at SemEval-2019 Task 5: Using Sentence embeddings to Identify Hate Speech Against Immigrants and Women in Twitter. SemEval 2019, p. 70-74.
[3] D. Reich, A. Todoki, R. Dowsley, M. De Cock, A. Nascimento. Privacy-Preserving Classification of Personal Text Messages with Secure Multi-Party Computation. NeurIPS2019

Bar-Ilan University
SCHOOL OF ENGINEERING & TECHNOLOGY
UNIVERSITY of WASHINGTON | TACOMA