Your Email Address Holds the Key:

Understanding the Connection Between Email and Password Security with Deep Learning

* Nina Mainusch
* Etienne Salimbeni
Dario Pasquini

Motivation

Passwords

are ubiquitous and vulnerable

They can depend on:

- users' language
- users' hobbies
- users' pets
- users' birthday
- etc.

Examples

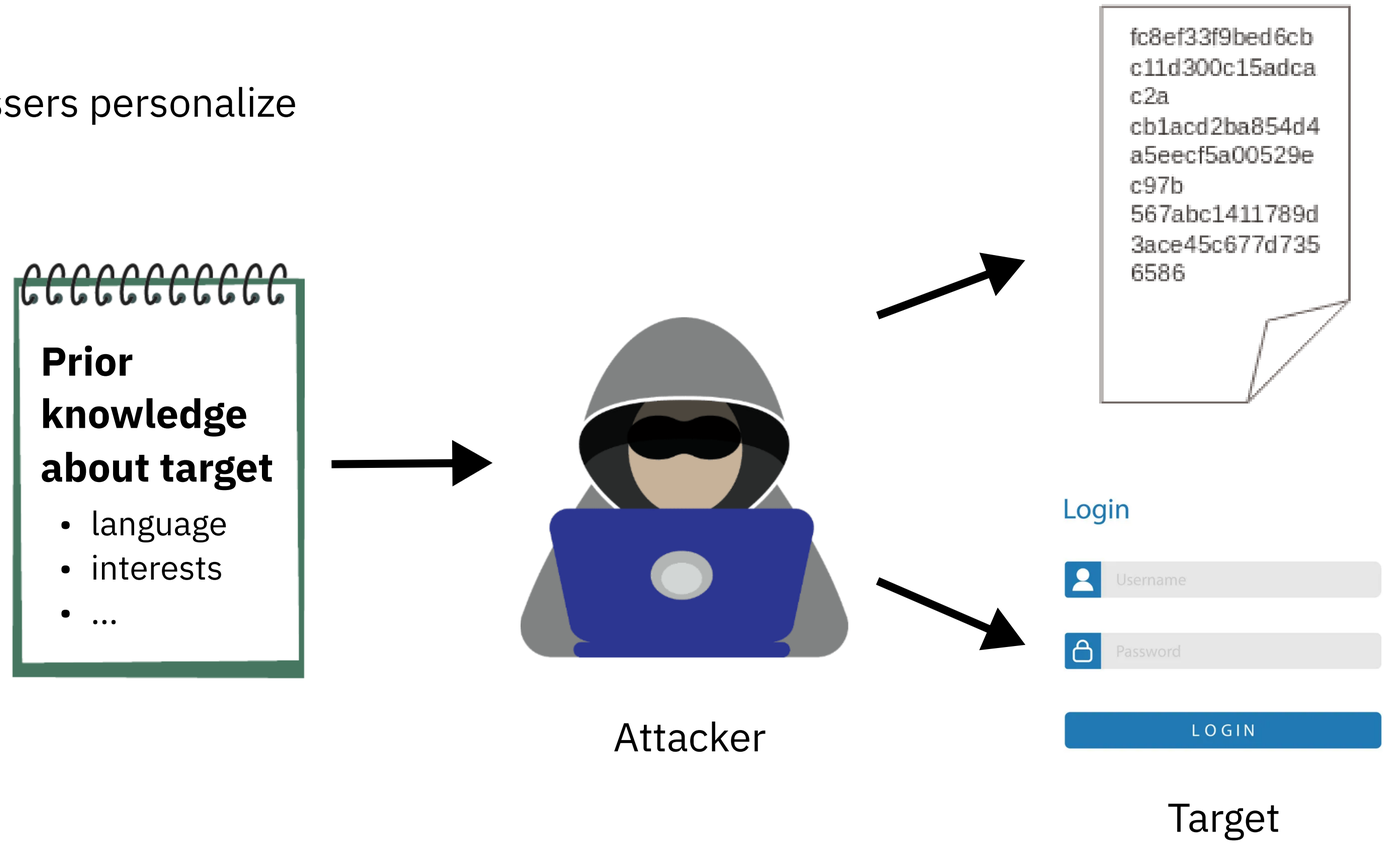
Email address of a user	user password
epowka@mail.ru	**epowka**
chocolate87@live.**it**	**dolcecioccolata**
ne61vin80@epost.**uk**	**arsenal**soccer14
koe-**dog**.shopping@hotmail.com	**fluffy**3000

Motivation

Password

guessing in the real world

Real-world password guessers personalize their attack to their target.

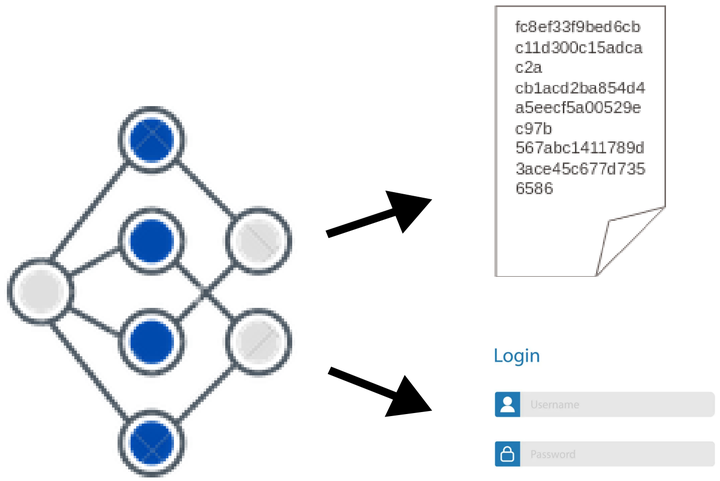


Motivation

Problematic

Current password attack models

- off-the-shelf rule-based models
- probabilistic models of a general password distribution



Attack model

Target

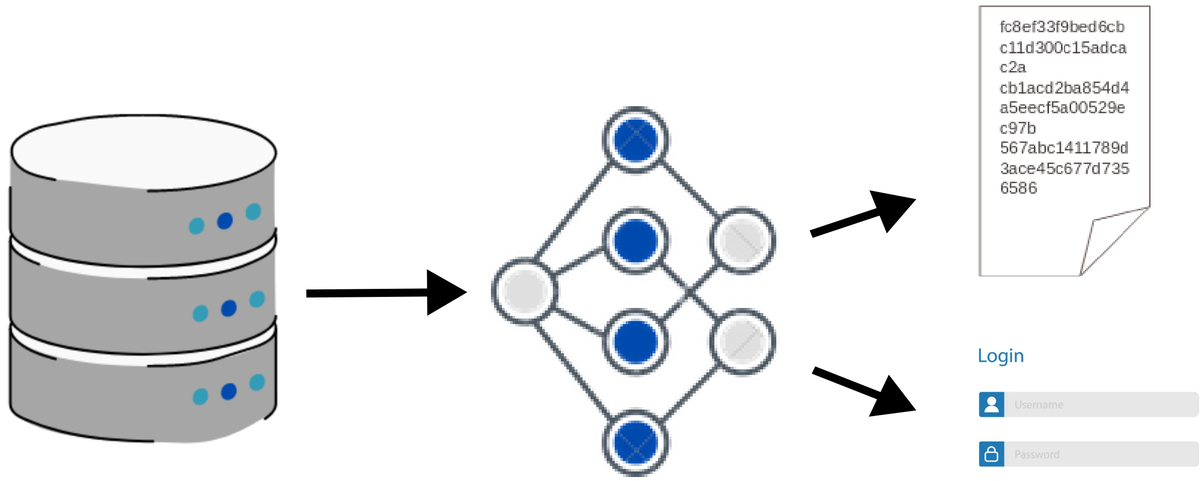
However

- Passwords are usually leaked with an email address
- Current models neglect this information

Motivation

Hypothesis

We approximate the attackers advantage with a context-aware deep learning model and a huge amount of data.



Prior knowledge
- Email address

Attack model

Target

We postulate that there is a correlation between email and password such that given a password `x` and an email address `email`:

$$P(x|email) > P(x)$$

Models & Methods

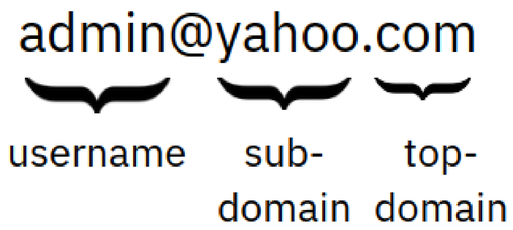
Methods

Dataset

Collection of password leaks

- We are using a collection of publicly available email-password pair leaks from Exploit.in and Anti Public dumps
- The dataset comprises **1.4 billion** entries

Processing the email



ilya_zhdanov@mail.ru	superilya
rca1@myspace.com	9694690k
escouser@gmail.com	kik192mbh
truonghvhh@yahoo.com	giang9a1
fatmoney_37@yahoo.com	money1
sam-i-am821@sbcglobal.net	twisted1
mdje@hotmail.co.uk	57trousers
alejandropache0714@yahoo.com	alcon1285
by_sanalcocuq@windowslive.com	nsh6am9ill
rlindner01@gmail.com	memories
drkemalpelit@hotmail.com	53495349
buttitshot@yahoo.com	monkey2
macomejestai@hotmail.it	ricciardo1977
p.dedek@seznam.cz	karelyfe
chinarmer@sbcglobal.net	cjs2cjay
tiphany75@live.fr	jeremy75
aldomolina_1996@hotmail.com	pumas1954
hechizada36@hotmail.com	amelie
seorsain@yahoo.com	littlegeorge
tri.at@hotmail.it	vaffanculo
shaterria@myway.com	beverly1
dudchenko-sergey@list.ru	14031986
moskotina74@inbox.ru	novosib
guchiwa@hotmail.fr	dosewisi
vermenyet@freemail.hu	Sari0320!
al208@yahoo.com	alexey
sebaporto@hotmail.com	07cf04cb

Related Work

Base model

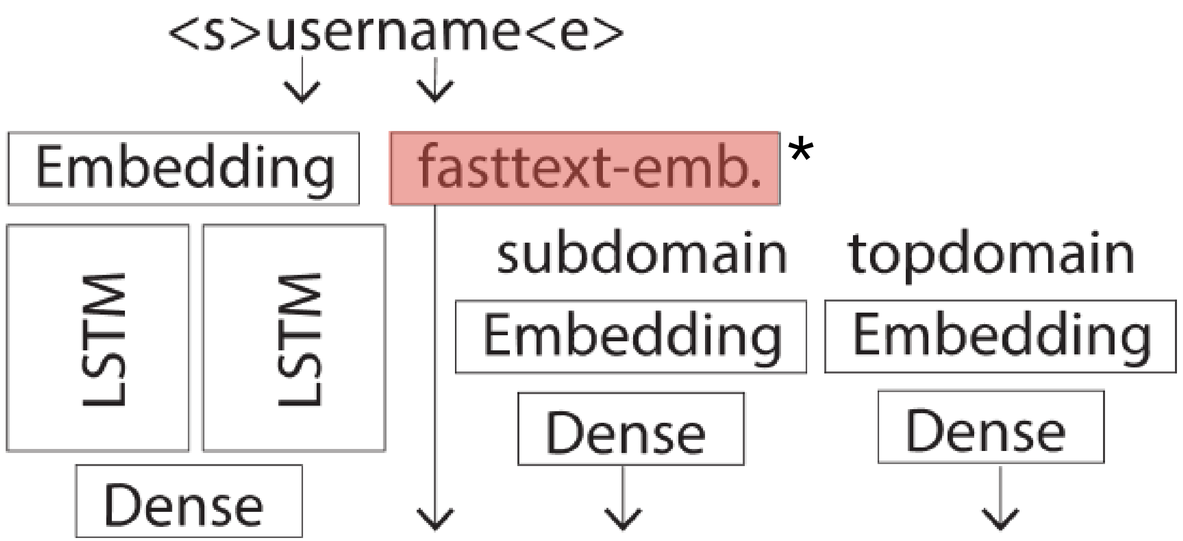
password-to-password

This architecture was originally proposed by:

W. Melicher, B. Ur, S. M. Segreti, S. Komanduri, L. Bauer, N. Christin, and L. F. Cranor. Fast, lean, and accurate: Modeling password guessability using neural networks. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 175–191, Austin, TX, Aug. 2016. USENIX Association.

Methods

Context-aware model



*P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

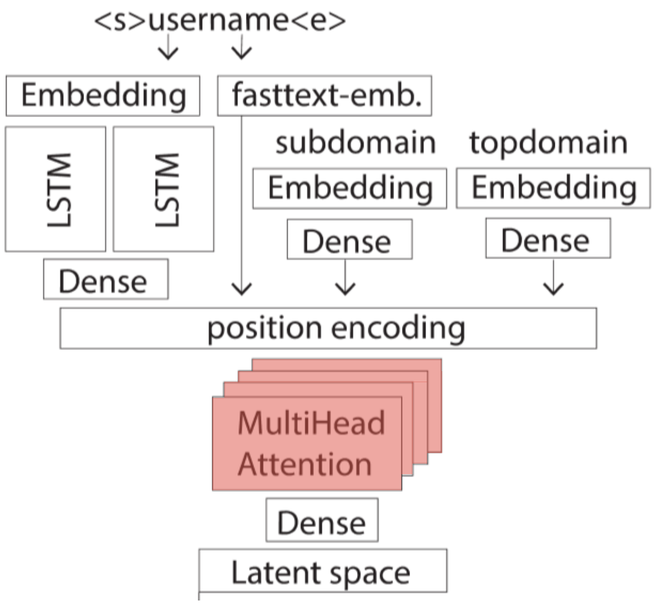
Fasttext model

- Semantic information in the username can be reflected by the password

→ improvement in test accuracy

Methods

Context-aware model



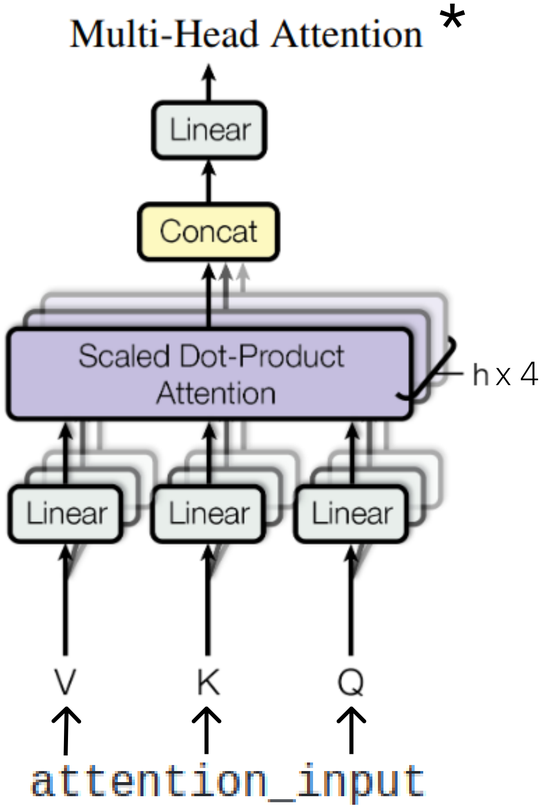
Multi-Head Self-Attention

- Enables us to investigate the importance of each embedding during inference

→ improvement in test accuracy

Methods

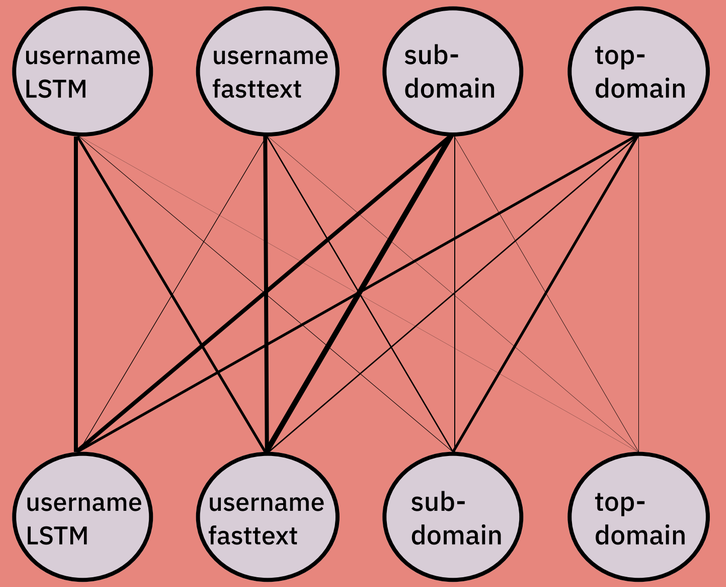
Context-aware model



[lstm-emb. username, fasttext-emb. username, emb. subdomain, embd. topdomain]

Multi-Head Self-Attention

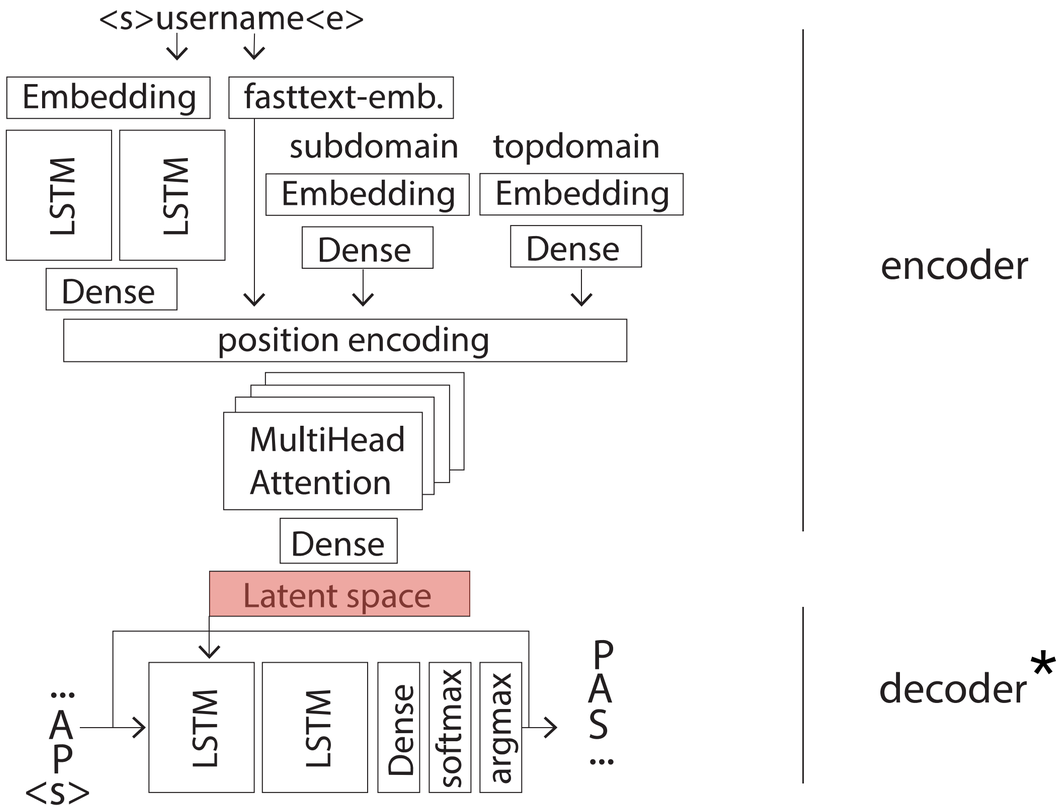
Each head learns its attention weights



* Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (p./pp. 5998--6008), .

Methods

Context-aware model



Latent space

- The latent space of the encoder serves as the seed of the decoder LSTM

* Same architecture as the base model

Evaluation Metric

Metric

Guess number

$$\mathcal{G}_{\mathcal{M}}(x) = |\{y \in \Gamma : \mathcal{M}(y) > \mathcal{M}(x)\}|$$

$\mathcal{M}$ = model

x = password

Γ = set of all allowed passwords

Examples

Email	Password	Guess number
donkacosta@hotmail.com	electrocolombo53	9e+08
celticbiker80@gmail.com	E$R%T6y7u8i(O)P	8e+23

Results
Discussion

Results

General Inference

Over the 500 000 samples of the test set

$$\mathcal{P}_{c-s} > \mathcal{P}_{base} \approx 60\% \text{ of the cases.}$$

Results

Base Model

email	pwd	$\mathcal{P}_{ctx}$	$\mathcal{P}_{base}$
kara_kartal541@hotmail.com	12345	**0.018454**	0.018093
travist_yudh@yahoo.com	12345	**0.018592**	0.018093
loveyougodpidra@yahoo.com	12345	**0.020903**	0.018093

Pairs with highest probabilities

capture most used password in the dataset

email	pwd	$\mathcal{P}_{ctx}$	$\mathcal{P}_{base}$
it@capnajax.com	iB]Wzx;V3u3%cw¿w)¡Mz2=3QPJW479	2.95e-78	**2.92e-75**
yrm2e-lkgjq-2x87v-2idrl-3myeq@live.com	YRM2E-LKGJQ-2X87V-2IDRL-3MYEQ	**7.79e-59**	1.18e-59
w@razryv.1gb.ru	w@razryv.1gb.ruw@razryv.1gb.ru	**2.07e-54**	1.73e-55

Pairs with lowest probabilities

long and random like passwords

Results

Context Aware Model

email	pwd	$\mathcal{P}_{ctx}$	$\mathcal{P}_{base}$
wangzi160@tianya.cn	123456	0.087573	**0.01577**
qingtaowang2003@tianya.cn	123456	0.085443	**0.01577**
sunguirong1234@tianya.cn	123456	0.085423	**0.01577**

Pairs with highest probabilities

capture most used passwords per group of emails

email	pwd	$\mathcal{P}_{ctx}$	$\mathcal{P}_{base}$
it@capnajax.com	iB]Wzx;V3u3%cw¿w)¡Mz2=3QPJW479	2.95e-78	**2.92e-75**
yrm2e-lkgjq-2x87v-2idrl-3myeq@live.com	YRM2E-LKGJQ-2X87V-2IDRL-3MYEQ	**7.79e-59**	1.18e-59
www.frank_ungemach@web.de	bc?watfngjl&-wvs-qmv-zv$v	1.64e-56	**3.03e-52**
whiskas-forever2008@yandex.ru	F8910fzZkvirus005Lerik@vIel	1.68e-54	**1.51e-53**

Pairs with lowest probabilities

long and random like passwords

Results

Best difference score

Similar email - password

Results

Best difference score

email	pwd	$\log(G_{base}/G_{ctx})$
celticbiker80@gmail.com	#E$R%T6y7u8i(O)P	25

Results

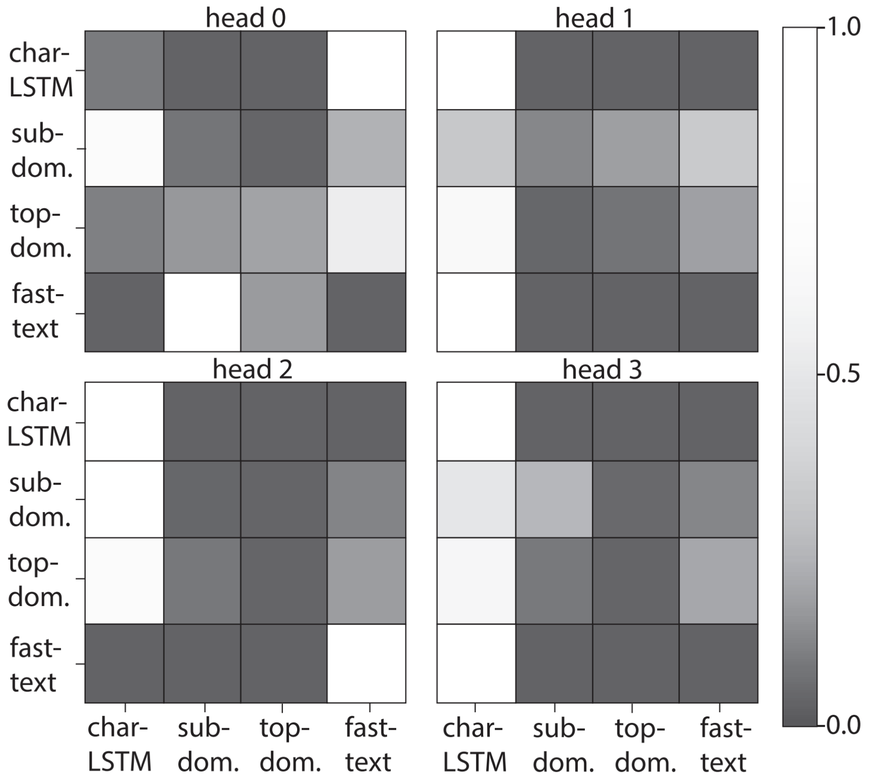
Worst difference score

email	pwd	$\log(G_{base}/G_{ctx})$
daleek1@lyme.in	d9Zufqd92N	-5
francisqg20@kamryn.tia.inxes.in	d9Zufqd92N	-5
biancabt4@mail.oplog.in	d9Zufqd92N	-6
junevb16@imp.chi.blognet.in	d9Zufqd92N	-4

Same passwords

Results

Attention matrix scores

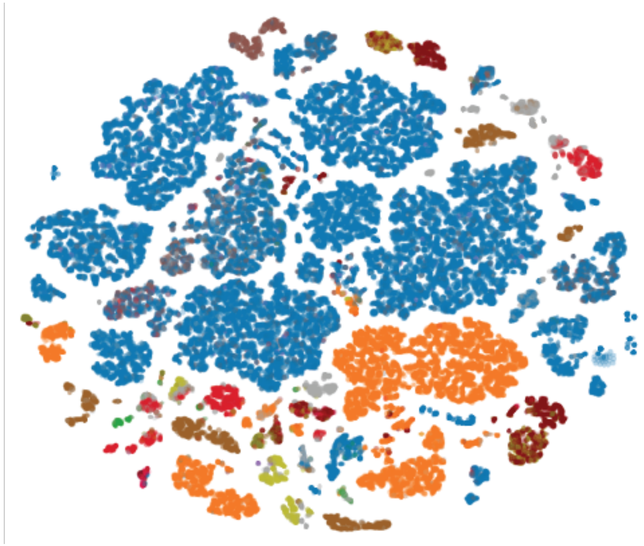


	username LSTM	sub domain	top domain	username Fasttext
cumulative attention score	**3.41**	2.64	2.91	3.27

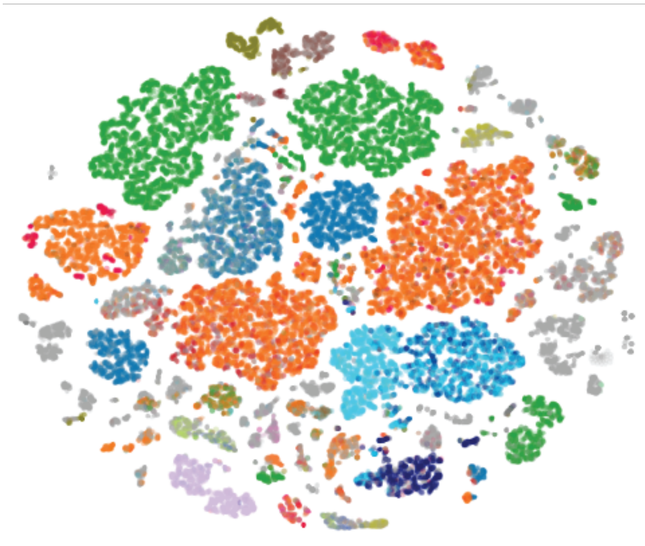
Results

Latent space exploration

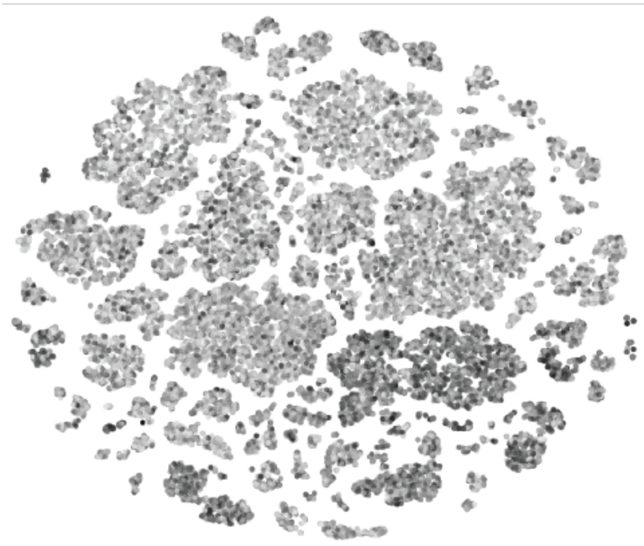
TSNE dimentionality reduction



Color by topdomain



Color by subdomain



Color by probability difference between models, the darker the better the context model is.

Conclusion

Conclusion

- **current simulations give a false sense of security**
- **context-aware models find sub-population more vulnerable to password guessing**
- **context-aware models are better at simulating a real-world attack**

- **Think twice how to choose your**

pA33w0rd!
PASSWORD
pppasswordd
password