eDarkFind: Unsupervised Multi-view Learning for Sybil Account Detection

Ramnath Kumar, Shweta Yadav, Raminta Daniulaityte, Francois Lamy, Krishnaprasad Thirunarayan, Usha Lokala, Amit Sheth

BITS Pilani Hyderabad Campus

KNO·E·SIS

Mahidol University

WRIGHT STATE UNIVERSITY

Motivation

- Darknet markets have grown substantially even with government interventions from 2013-2016 [1]

Feature	Growth
Total revenue	2x
Total number of transactions	3x
Total number of listings	5.5x
Total number of listings per vendor	2x

Incremental growth of the Darknet Market [1]

[1] Kristy Kruithof. 2016. Internet-facilitated drugs trade: An analysis of the size, scope and the role of the Netherlands. RAND.

Motivation

- Trend in switching from DarkMarket to P2P markets which forces a new method of monitoring users. [2]

- Analysis of new drugs with the help of association with users (**comparative advantage**)

Vendor A	Cost of good	Opportunity Cost
Good x	X - 3$	X - 4$
Good y	Y - 4$	Y - 3$

Hence, vendor A should produce good Y.

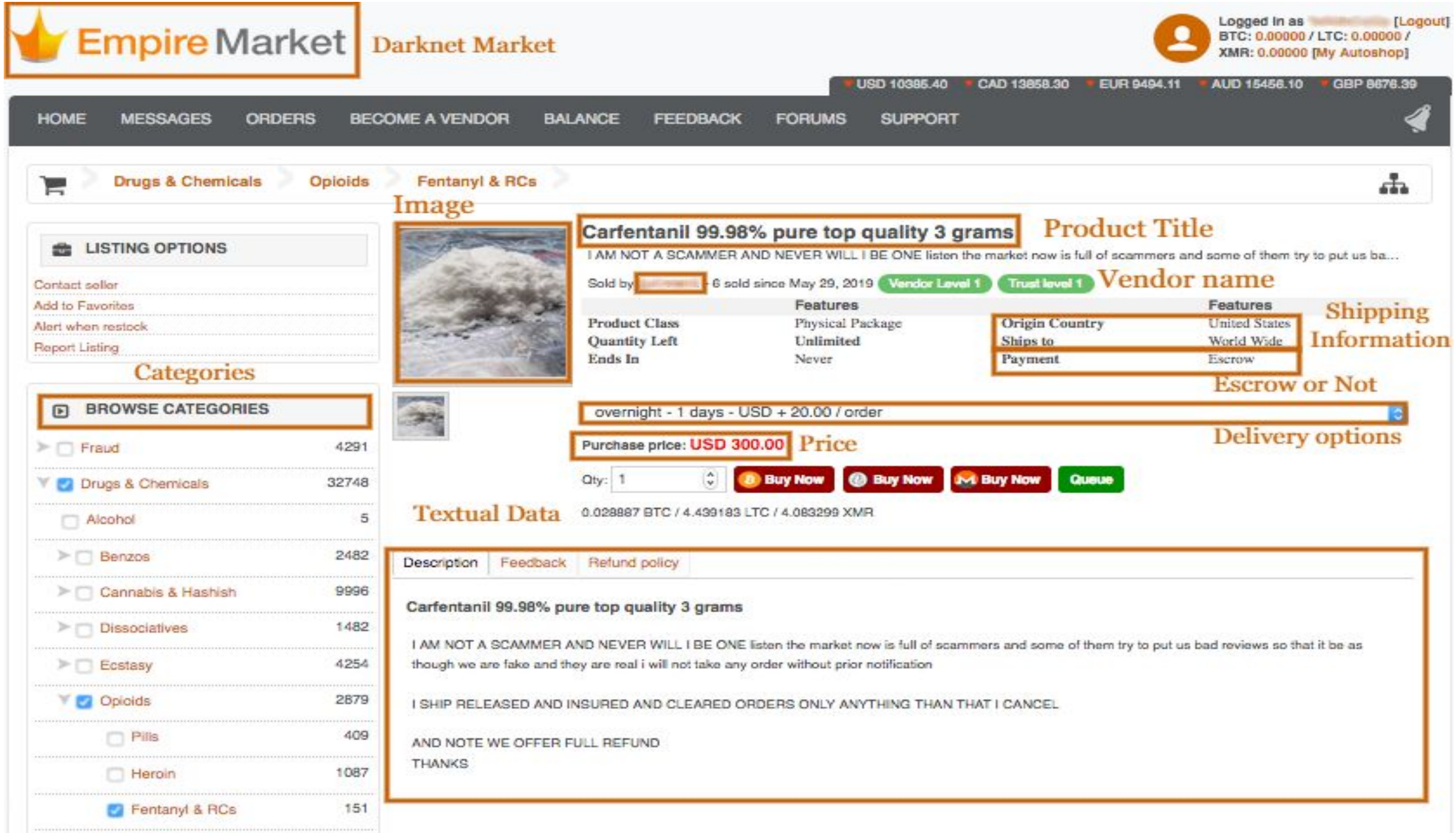
[2] https://thenextweb.com/insider/2017/10/23/dark-web-drug-vendors-p2p-shop/

Motivation

- Help to assess the real number of vendors on darknet markets
- Previous approaches rely heavily on supervised learning [3,4]

[3] Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. Your Style Your Identity: Leveraging Writing and Photography Styles for Drug Trafficker Identification in Darknet Markets over Attributed Heterogeneous Information Network. In The World Wide Web Conference. ACM, 3448–3454.
[4] Thanh Nghia Ho and Wee Keong Ng. 2016. Application of stylometry to dark-web forum user identification. In International Conference on Information and Communications Security. Springer, 173–183.

Data Available



Snapshot of Darknet Market

Problem Statement

- Identification of sybil accounts on the dark web
- Define user similarity between vendors by making use of various views (**Vendor Embedding**)
- Vendor embedding formed using:
    - Textual Data
    - Substance information
    - Location Information

Dataset

- Data extracted using eDarkTrends [5]
- 1992 unique vendors collected over 3 different sites.
- Extracted textual data, location and substance information
- DAO ontology [6] used in this process to capture slangs, route of administration, etc.

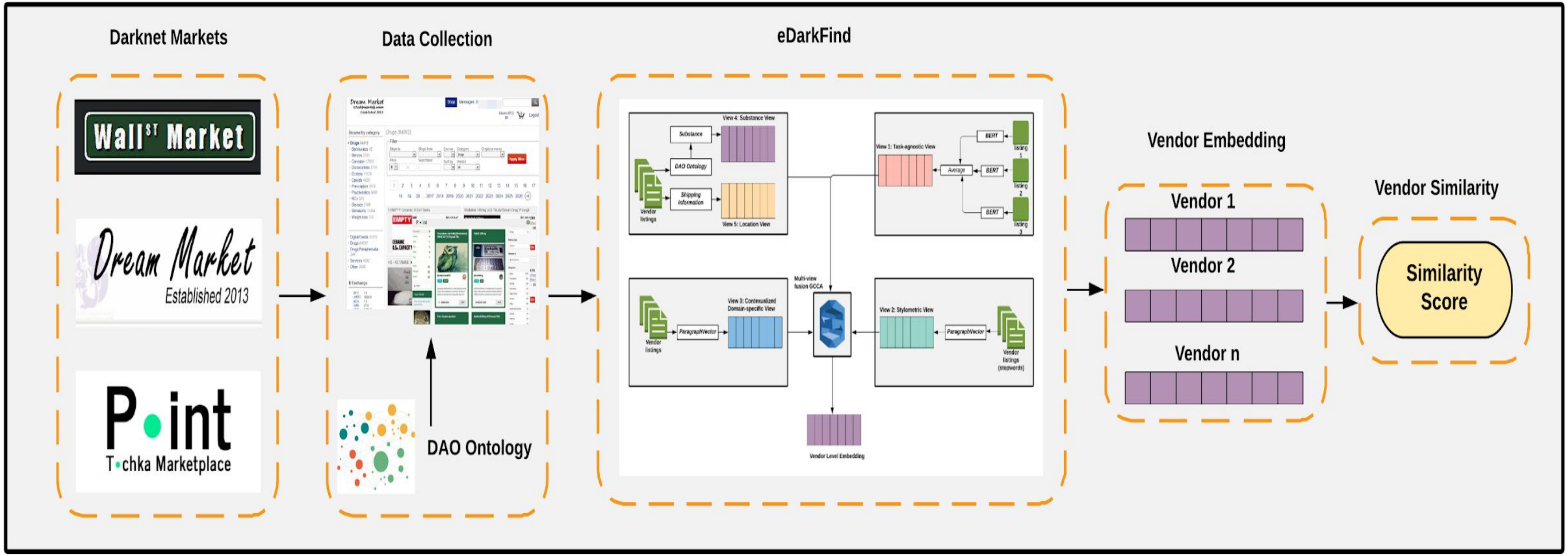
[5] Usha Lokala, Francois R Lamy, Raminta Daniulaityte, Amit Sheth, Ramzi W Nahhas, Jason I Roden, Shweta Yadav, and Robert G Carlson. 2019. Global trends, local harms: availability of fentanyl-type drugs on the dark web and accidental overdoses in Ohio. Computational and Mathematical Organization Theory 25, 1 (2019), 48–59.
[6] Cameron, Delroy, et al. "PREDOSE: a semantic web platform for drug abuse epidemiology using social media." Journal of biomedical informatics 46.6 (2013): 985-997.

Dataset

	Dream Market	Tochka	Wall street	All
Unique # Vendor names	1448	408	466	1992
Unique # Substance	852	313	290	1148
Unique # Location	356	44	29	389
Unique # Descriptions	16800	1829	1723	18472

Summary of Dataset

Methodology



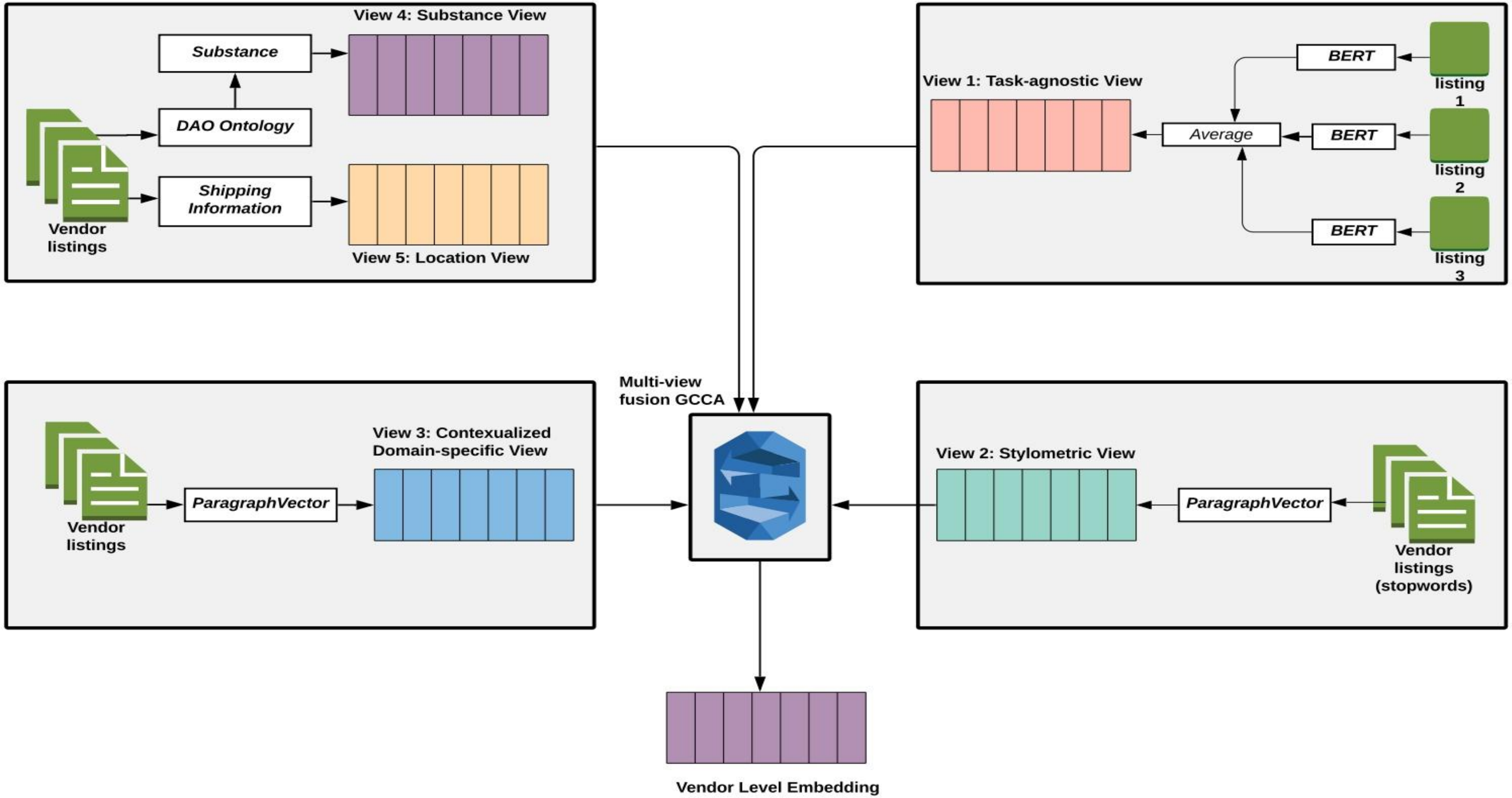
Overview of the proposed model architecture

Multi-view Learning

- Multi-view learning is an ideal learning mechanism for the data where examples are characterized by distinct (often orthogonal) feature sets (views).

- Allows us to capture vendor embedding, which is better than capturing multiple views of the vendor.

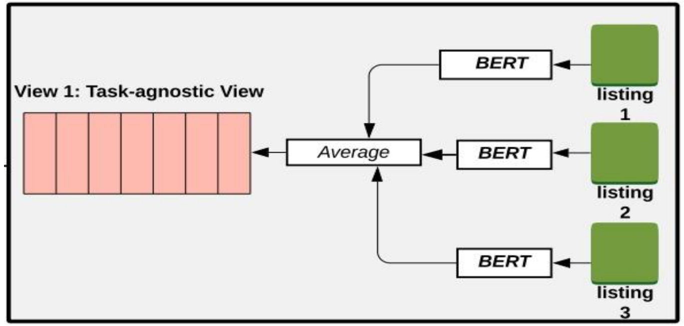
Summary of Approach



eDarkFind Model

Task Agnostic View

- To capture the semantics behind the textual data posted by the vendor on generic corpus
- We used Bidirectional Encoder Representations from Transformers (BERT) [7]

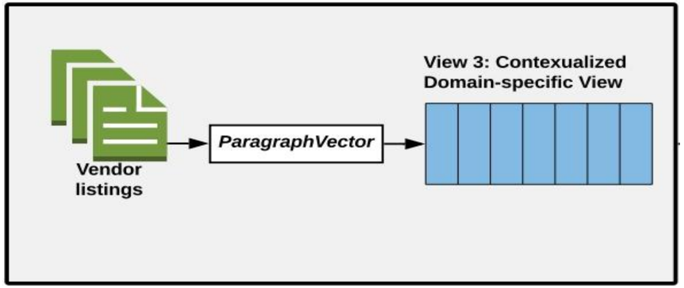


Task Agnostic View

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).

Contextualized Domain-Specific View

- To capture the semantics behind the textual data posted by the vendor on domain specific corpus
- Trained the vector using ParagraphVector[8] model

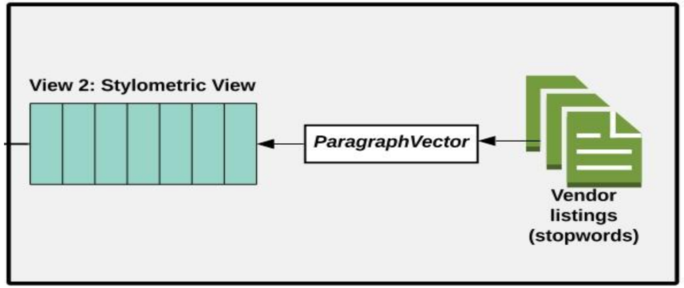


Contextualized Domain-Specific View

[8] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International conference on machine learning. 1188–1196.

Stylometric View

- To capture the style of writing of the vendor.
- Trained the vector using ParagraphVector[8] model
- Applied on only stopwords and special characters



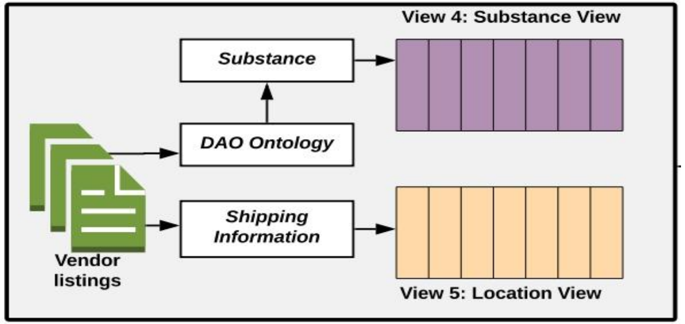
Stylometric View

[8] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In International conference on machine learning. 1188–1196.

Location and Substance View

- Capture data from location and substance fields

- Use of alternate and slang terms. Eg. Suomi for Finland



Stylometric View

Location and Substance View

- Use simple binary embedding:

  eg.

USA	CAN	ESP	IND	CHN	BEL	NOR	NZL	SAU	UKR
1	1	0	0	0	0	0	0	0	0

- Add a self information weight or information content, for all features

  Information content = $w_i = -log(Pr(F_i^{all} = 1))$

Fusion

- Cannot simply concatenate since each vector may correspond to different modalities (image vs text) or very different distributional properties

- These views are fused using CCA [9] to obtain a single representation, which we call Vendor embedding

- Allows us to infer information from cross variance matrices

- Employ an extension called weighted generalized CCA.

[9] Harold Hotelling. 1992. Relations between two sets of variates. In Breakthroughs in statistics. Springer, 162–190.

Experiments

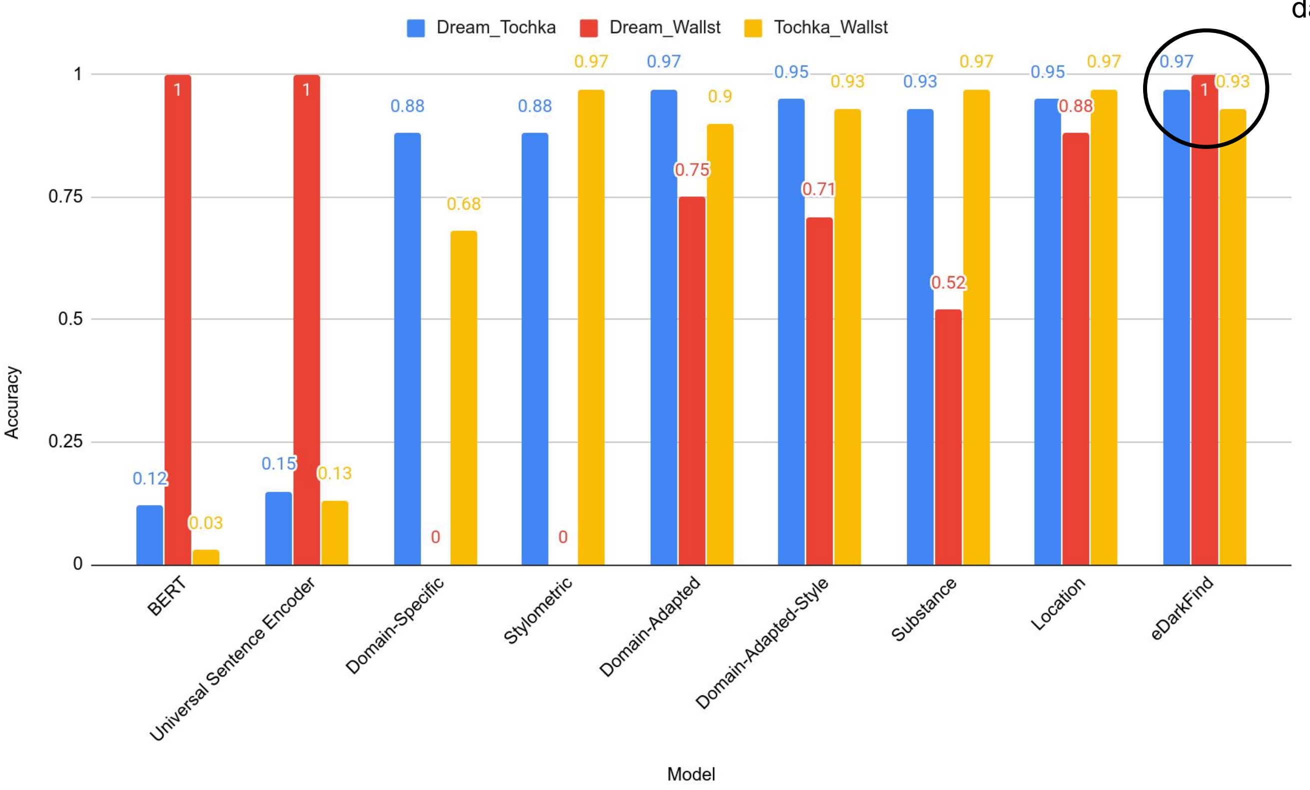
- <V1, V2> -> S
  - V1, V2 : vendors
  - S : target variable
- Created 3 cross domain datasets:
  - Dream_Tochka
  - Dream_Wallst
  - Tochka_Wallst

Experiments

- Compute similarity score and used threshold of 0.5
- Baselines include:
  - BERT
  - Universal Sentence Encoder
  - Domain Specific
  - Stylometric
  - Domain Adapted
  - Domain Adapted with Style
  - Substance
  - Location
  - eDarkFind

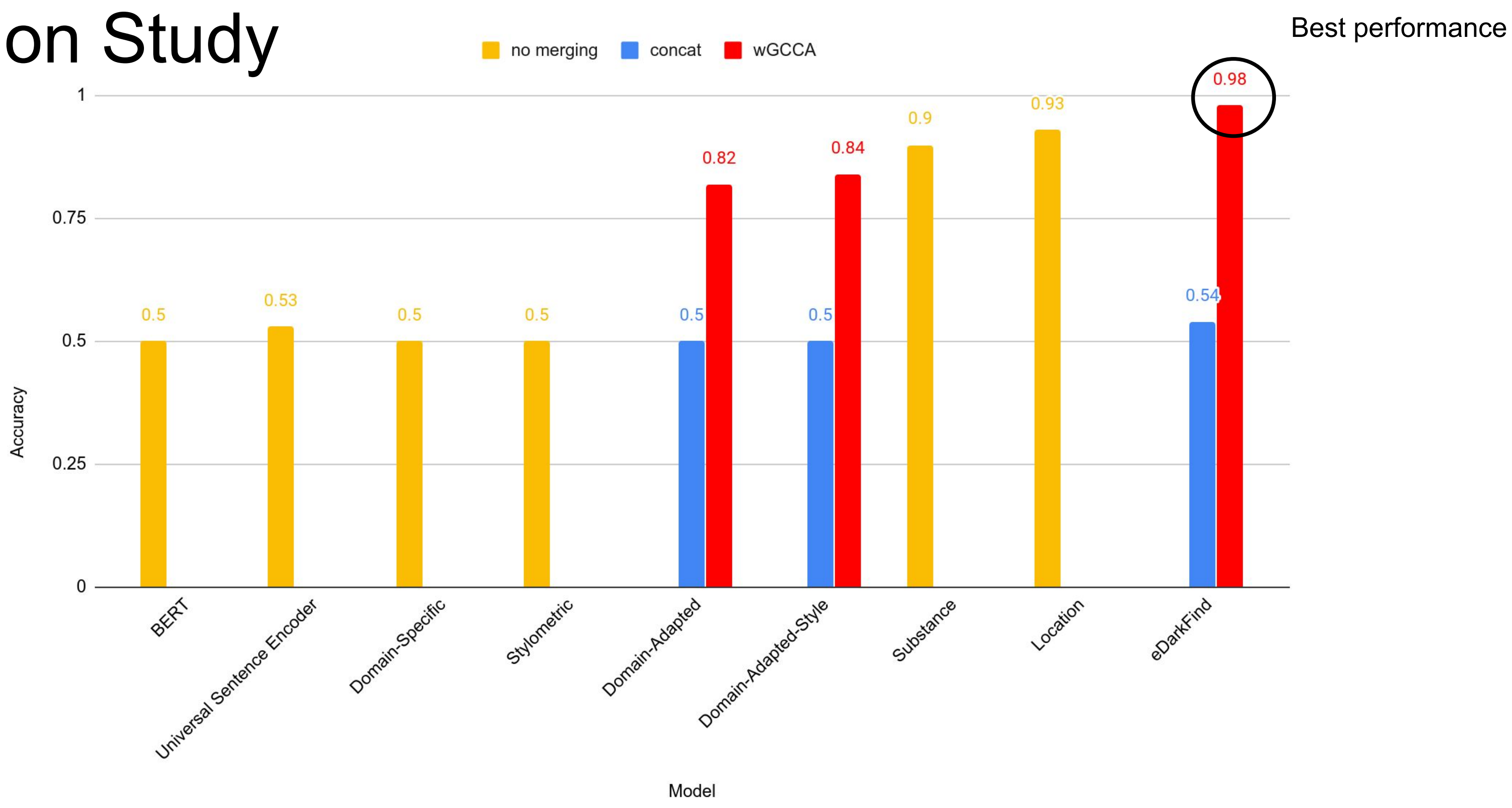
Results

Highest average accuracy across all datasets



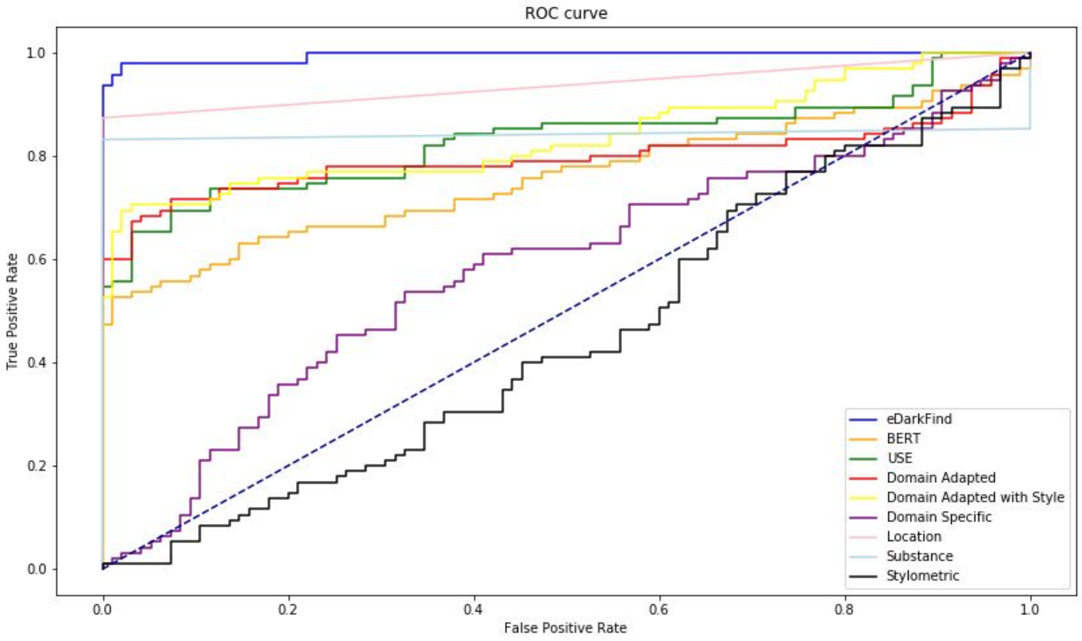
Performance metric of our model on different datasets

Ablation Study



Performance metric of various models on All sites combined.

Results



ROC curve comparison between true positive rate & false positive rate over the baselines and proposed models

Domain Specific Error Analysis

- Multilingual Data
- Use slang terms across listings captured by our model (e.g., horse for heroin)
- Lack of uniform features in website adds noise to our model. (product description and rating data)
- Some vendors may operate from different locations or may even be selling different drugs
- Branding is common in these markets

Case Studies	@Vendor 1	@Vendor 2
Branding	5//02/14 09:49 am,5/Thanks alles schick/11/10 01:46 pm, <END>Tilidin 50MG/4MGOriginal Apothekenware	5//02/14 09:49 am,5/Thanks alles schick/11/10 01:46 pm, <END>Tilidin 50MG/4MGOriginal Apothekenware <END> 5/Thanks alles schick/11/10 01:46 pm,
Comparing product Description and rating since the vendor did not enter product description in other site.	Percocet Oxycodone 5/325mg 200 TabletsFinalize Early and get 20 Free bonus sent for a total of 220!US Made Mallinckrodt 5mg/325 (made in St. Louis, Miss. USA) ...	5//02/07 01:03 pm,5/Thanks Again. A++/01/21 11:49 pm,5/Trustworthy/01/16 12:22 pm,4.33//01/07 08:50 am,5/Great communication, trustworthy, and overdelivered./12/31 11:09 pm,5//11/29 03:25 pm,5/FAST A+++ Best Stealth I've seen yet.
Similar stylometric Features captured by the use of special characters or emojis.	****** NEWS 25.12.2018 NEWS ****** We ship all new ...	PRODUCTS AFGHAN HEROIN A+++COCAINE #3 ...

Use Case examples

Conclusion

- Multi-view learning Sybil account detection on the real-life Darknet market dataset achieving an accuracy of 98%

- Performed cross-domain analysis to justify uniform results

- Explored domain specific knowledge graph of drug (DAO) in sybil account detection

Thanks!

Any questions ?

You can find me at

- ramnathkumar181@gmail.com