Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification Task

UO-UPV

Roberto Labadie Tamayo [1] Daniel Castro Castro [1] Reynier Ortega Bueno [2]

[1] Universidad de Oriente, Cuba
[2] PRHLT Research Center, Universitat Politècnica de València

September, 2021

Outline

Task and Datasets

Our Approach

Experiments

Conclusions and Future Work

Source Code: `https://github.com/labadier/hatespeechspread`

Problem Statement

Given

$$\mathcal{D} = \{(X_i, y_i)\}_{i=1}^{n}$$
$$X_i = \{x_{ij}\} \quad j = 1...200, \text{ and } x_{ij} \in \mathbb{W}, y_i \in \{0, 1\}$$

Where

- X_i is a set of tweets belonging to the i^{th} user
- y_i represents whether user i is Hate Speech Spreader or not
- $\mathbb{W}^*$ is the set of all possible strings

Find

$$\mathcal{F} : S \rightarrow \{0, 1\}$$

Dataset

Profiling Hate Speech Spreaders Dataset

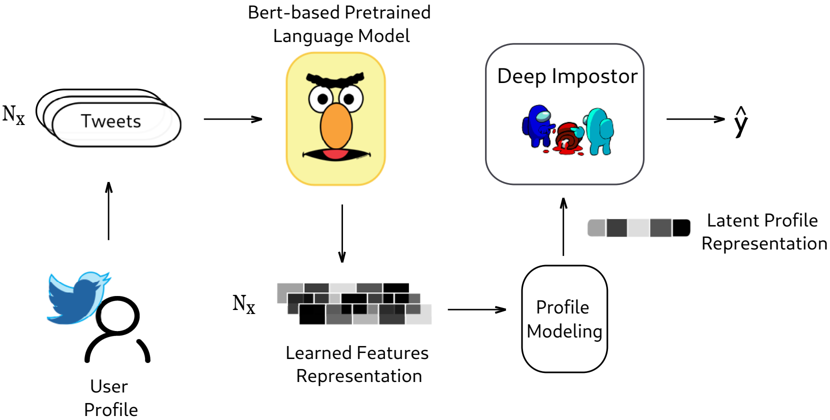
Class	Language	
	EN	*ES*
Hate Spreader	100	100
No Hate Spreader	100	100

SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter Dataset

Class	Language	
	EN	*ES*
Hateful	3783	1857
No Hateful	5217	2643

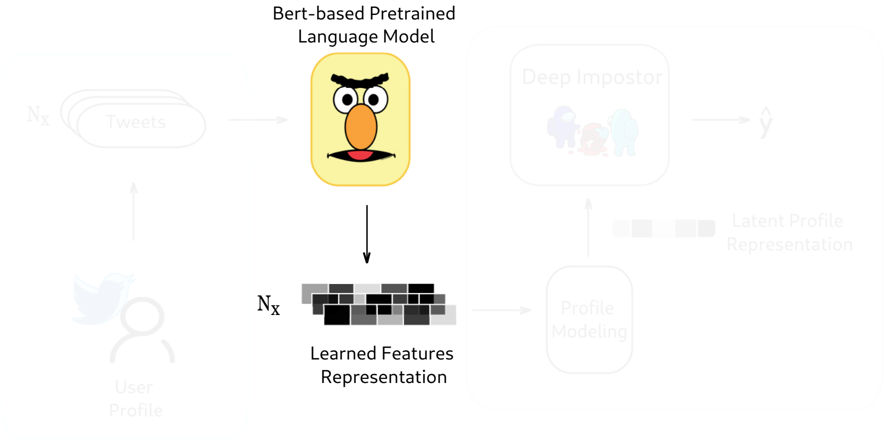
Overall Model

Modular Architecture for modeling and classifying user's profiles:


Bert-based Pretrained Language Model
Deep Impostor
N_x
Tweets
$\hat{y}$
Latent Profile Representation
N_x
Learned Features Representation
Profile Modeling
User Profile

Pretrained Language Models

Modular Architecture for modeling and classifying user's profiles:



Bert-based Pretrained Language Model

N_x

Tweets

User Profile

N_x

Learned Features Representation

Profile Modeling

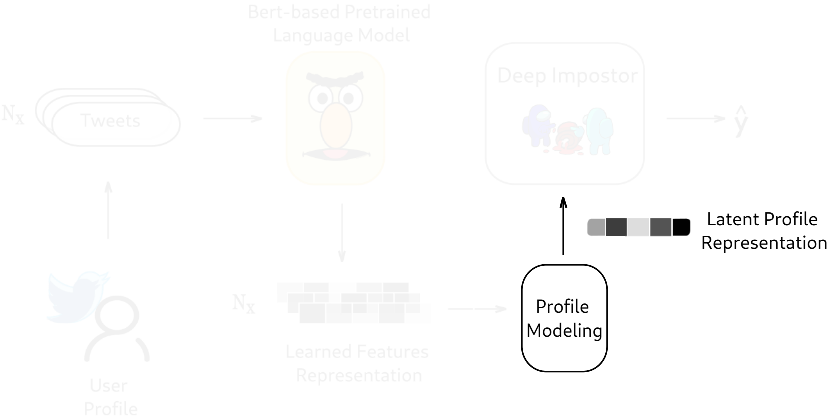
Latent Profile Representation

Deep Impostor

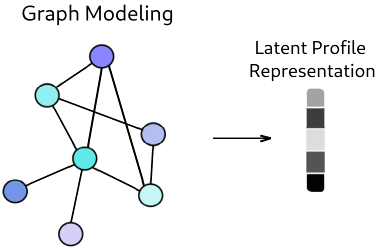
$\hat{y}$

Profile Modeling

Modular Architecture for modeling and classifying user's profiles:


Bert-based Pretrained Language Model
Deep Impostor
N_x
Tweets
$\hat{y}$
Latent Profile Representation
N_x
Profile Modeling
Learned Features Representation
User Profile

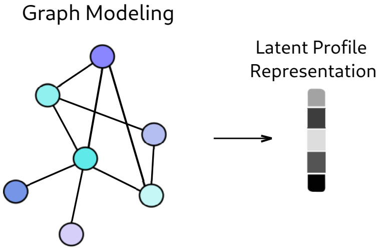
Graph-Based Profile Modeling



Graph Modeling

Latent Profile Representation

Graph-Based Profile Modeling



Graph Modeling

Latent Profile Representation

We employ the convolution operator defined as[1]:

$$X' = ReLU(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X\Theta)$$

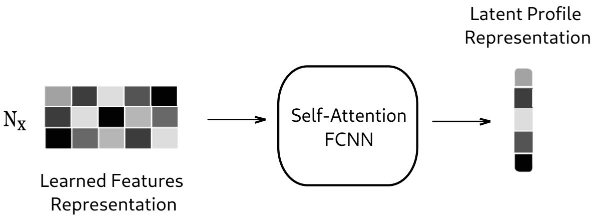
Node-wise Notation:

$$x_i' = ReLU\left(\Theta \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{1}{\sqrt{\hat{d}_j \hat{d}_i}} x_j\right)$$

[1] [Kipf et al. 2017. Semi-Supervised Classification with Graph Convolutional Networks]

Sequence-Based Profile Modeling

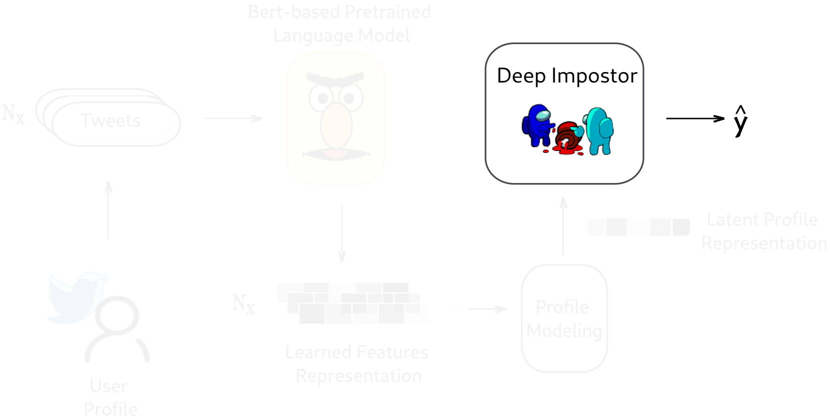
All elements are related as a sequential data but regardless any order.


Latent Profile Representation
N_x
Self-Attention FCNN
Learned Features Representation

[G. Zheng et al. 2018. OpenTag: Open Attribute Value Extraction from Product Profiles]

Deep Impostor Method

Modular Architecture for modeling and classifying user's profiles:



Bert-based Pretrained Language Model
N_x
Tweets
Deep Impostor
$\hat{y}$
Latent Profile Representation
N_x
Learned Features Representation
Profile Modeling
User Profile

[S. Seidman. 2013. Authorship Verification Using the Impostors Method]

Deep Impostor Method (DIM)

Let H and K be the sets of Hate Spreaders and No Hate Spreaders respectively and u an unknown profile, $\bar{H}$ and $\bar{K}$ are the randomly sampled prototypes from H and K.
Let $\mathcal{F}$ be a similarity function:

$$P_i(u, \bar{H}_i) = \begin{cases} 1 & \text{if } \sum_{j}^{|\bar{K}_i|} [\mathcal{F}(u, \bar{H}_i) > \mathcal{F}(u, \bar{K}_{ij})] > \frac{|\bar{K}_i|}{2} \\ 0 & \text{otherwise} \end{cases}$$

Then, avoiding the feature selection phase:

$$\hat{y}(u) = \begin{cases} 1 & \text{if } \sum_{i}^{|\bar{H}|} P_i(u, \bar{H}_i) > \frac{|\bar{H}|}{2} \\ 0 & \text{otherwise} \end{cases}$$

Experiments

Impact of the Profile Modeling modules on the Profiling Hate Speech Spreader on Twitter task

Data	Language	Deep Model			
		SGCN-2	SGCN-3	Att-FCNN	Att-BiLSTM[2]
CV	English	0.76	0.76	0.75	0.77
	Spanish	0.83	0.75	0.88	0.82
	AVG	0.795	0.755	0.815	0.795
Test	English	0.49	0.51	0.73	0.79
	Spanish	0.59	0.51	0.81	0.74
	AVG	0.54	0.51	0.77	0.765

[2] [Labadie et al. 2020. Fusing Stylistic Features with Deep-Learning Methods for Profiling Fake News Spreader]

Experiments

Deep Impostor Method Performance

Data	Language	Profiling Model			
		SGCN	Att-FCNN	Att-BiLSTM[2]	AVG Method
CV	English	0.73	0.72	0.74	0.73
	Spanish	0.76	0.76	0.82	0.78
	AVG	0.745	0.74	0.78	0.755
Test	English	0.72	0.73	0.73	**0.74**
	Spanish	0.80	**0.85**	0.79	0.82
	AVG	0.76	**0.79**	0.76	**0.78**

[2] [Labadie et al. 2020. Fusing Stylistic Features with Deep-Learning Methods for Profiling Fake News Spreader]

Conclusions

- SGNN was not able to generalize well on the test data.
- Our adaptation of the Impostor Method outperformed the accuracy of DL methods
- Even when the performance of SGNN was not the expected on the test dataset, the DIM achieved encouraging results.
- The Attention-FCNN based representation obtained the best result through the DIM.

Future Work

- We plan exploring a Metric Learning Approach as $\mathcal{F}$ function for the DIM. Requiring more data.
- Exploring a more sophisticated prototype sampling technique, which involves similarity relations within the data, rather than sample randomly prototypes.
- Expressing the graph-based modeling through a more restrictive connection among the nodes, rather than connect them each other, and/or study an attention based aggregation function for message passing.

Deep Modeling of Latent Representations for Twitter Profiles on Hate Speech Spreaders Identification Task

UO-UPV

Roberto Labadie Tamayo [1] Daniel Castro Castro [1] Reynier Ortega Bueno [2]

[1] Universidad de Oriente, Cuba
[2] PRHLT Research Center, Universitat Politècnica de València

September, 2021