Supervised Classification of Social Spammers using a Similarity-based Markov Random Field Approach

Nour El-Mawass, Paul Honeine, Laurent Vercouter
LITIS, Université & INSA de Rouen, Rouen, France

Introduction

Problem

Malicious use of online social networks (OSNs) has a detrimental effect on these platforms' security, usefulness, profitability and information veracity. The evolving nature of the spam phenomenon, causes the many proposed supervised classifiers to become obsolete.

Contributions

The present work models the spam detection problem as a classification problem where the goal is to assign a label (legitimate vs. malicious) to a given social account on Twitter.

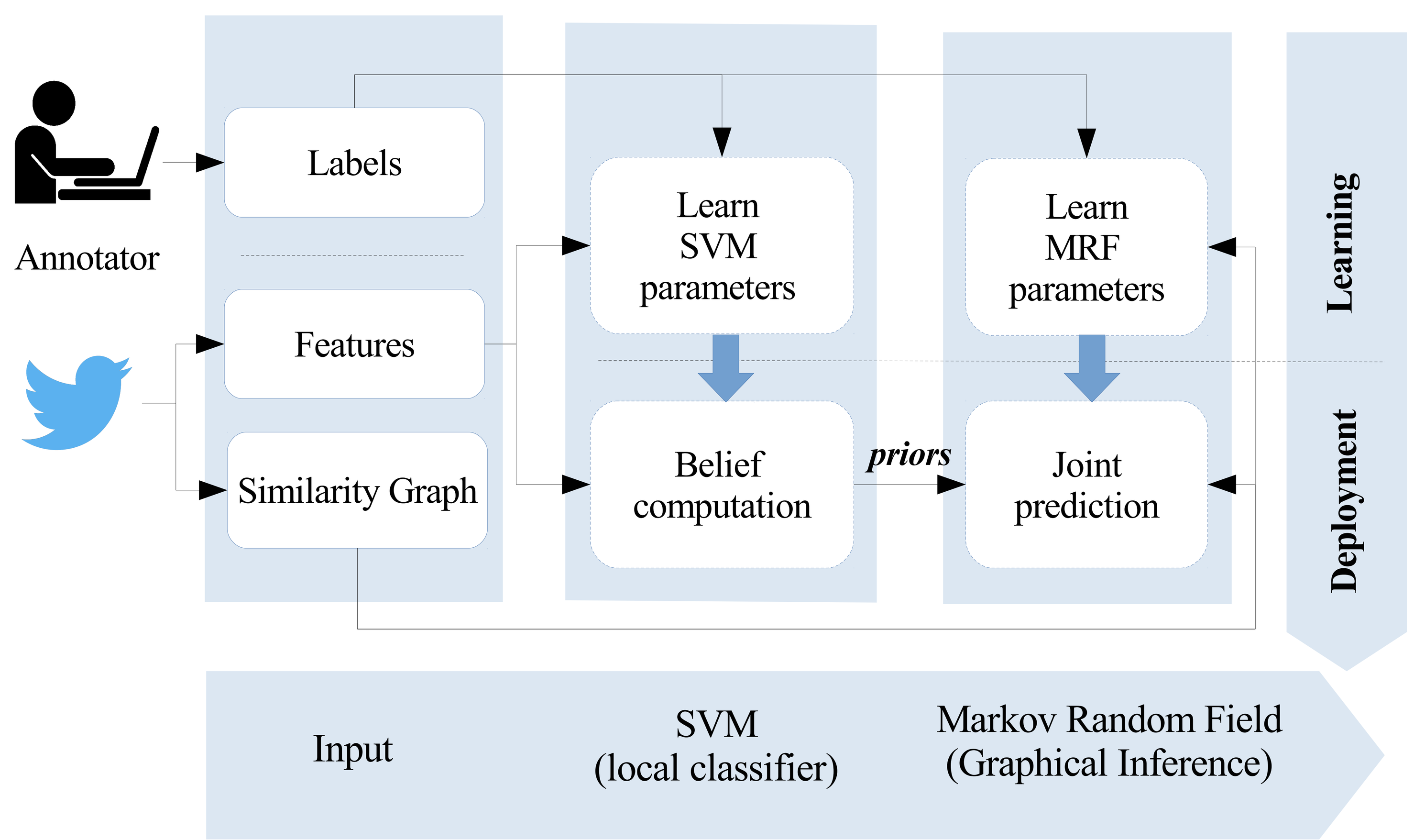
We propose to solve the problem by exploiting the similarity between social accounts and performing graphical inference over smilar accounts.

Dataset

- Data collected from Twitter (between 5 and 21 October 2017).
- A random sample of 20M tweets from 12M active users (+ tweets of selected users in the sample).
- Groundtruth dataset of 767 Twitter users labeled (mostly manually) as legitimate or spammers.

Proposed System

The proposed system leverages similarity between users to propagate and correct beliefs about their labels. We initiate beliefs using supervised classifiers trained with selected state-of-the-art features. These beliefs are subsequently used as node priors in the Markov Random Field (MRF). We apply joint optimization using Loopy Belief Propagation over the MRF to get the most probable configuration of labels.



Features

Three sets of features are compared:

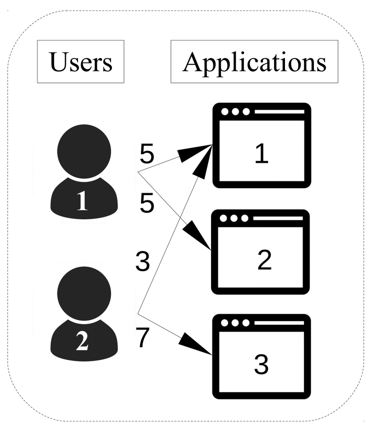
- Two sets of features proposed in prominent previous works (denoted here as Benevenuto [1] and Stringhini [2]).
- Our set of 28 state-of-the-art features, selected with domain knowledge and prioritized using information gain and Chi-squared selection criteria.

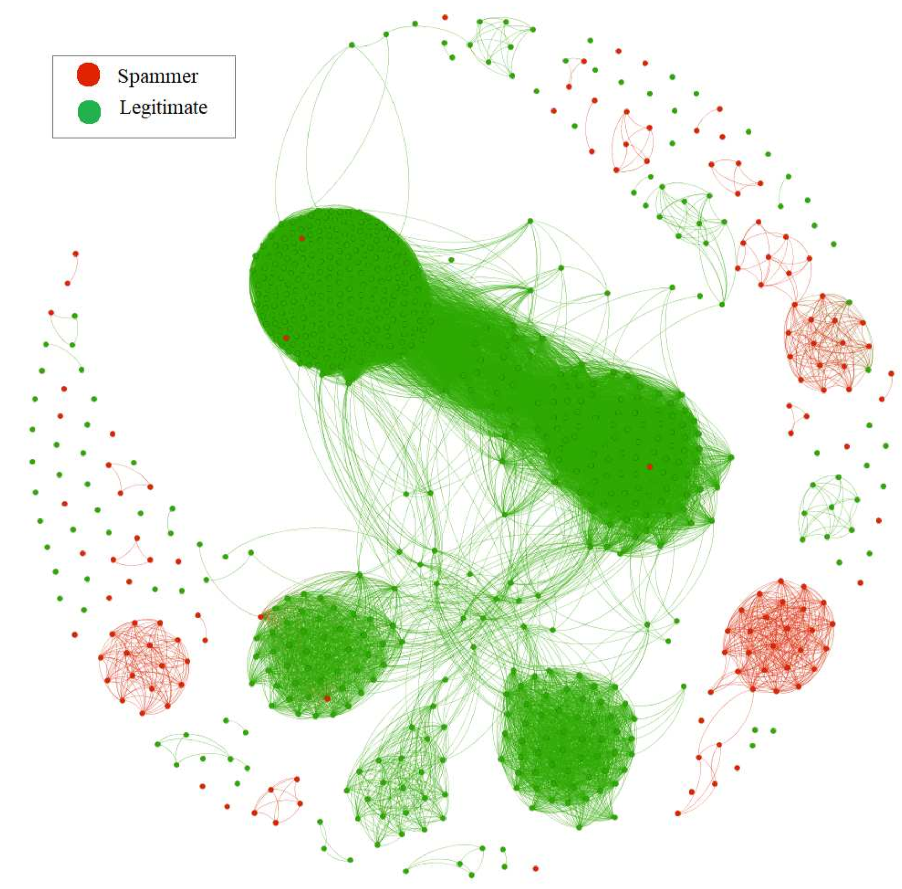
Table 1: The account features used to train local SVM classifier

Category	Features	Category	Features
Profile:	Age of the account Statuses count	Content:	Replicates Fraction of replies ⋮ *(+7 other features)*
Social net.:	Friends count Followers count ⋮ *(+2 other features)*	Behavior:	Avg intertweet interval Temporal distribution of tweets ⋮ *(+5 other features)*

Similarity Graph

The similarity graph between users is obtained with the cosine similarity between the applications profiles defined over each user.
The edge weight is equal to the similarity value.

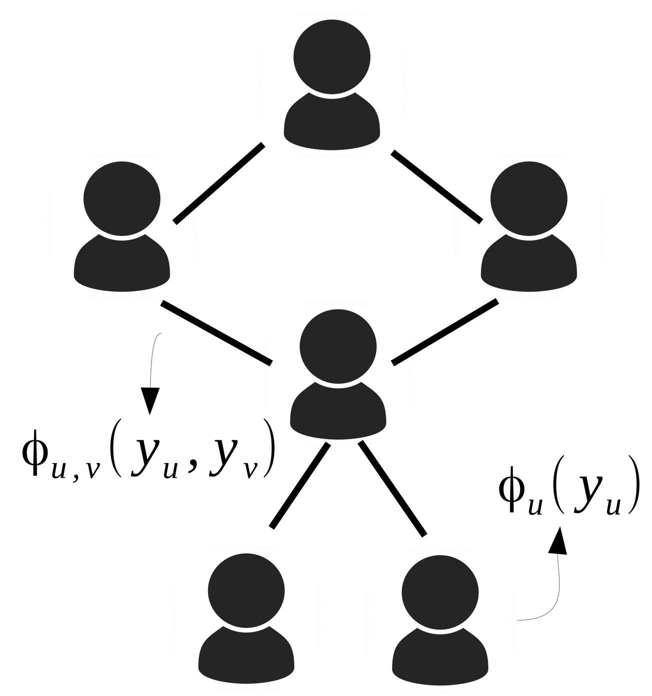




Markov Random Field

A Markov Random Field (MRF) is a probabilistic graphical model that allows joint inference over dependent random variables. It consists of a graph $G(V, E)$ where nodes are random variables and edges denote a dependency between two random variables. We use the pairwise MRF model (p-MRF), and define two types of potentials over nodes: edge potentials $\phi_{(u,v)}(Y_u, Y_v)$ and node potentials $\phi_v(Y_v)$. The goal is to maximize the probability of a joint configuration of labels $P(Y|\Theta)$ by optimizing the product of potentials:

$$P(Y|\Theta) = \frac{1}{Z}\prod_{v\in V}\phi_v(Y_v)\prod_{(u,v)\in E}\phi_{(u,v)}(Y_u, Y_v).$$



Experimental Evaluation & Results

Table 2: Classification results of SVM and our model on three different sets of features

		Our features		Benevenuto [1]		Stringhini [2]	
		Legitimate	Sybil	Legitimate	Sybil	Legitimate	Sybil
SVM	Precision	0.947	0.795	0.902	0.703	0.843	0.733
	Recall	0.952	0.778	0.941	0.578	0.978	0.244
	F-measure	0.949	0.787	0.921	0.634	0.905	0.367
	Accuracy	0.918		0.87		0.835	
SVM + MRF (this paper)	Precision	**0.968**	**0.878**	**0.91**	**0.774**	**0.877**	**0.8**
	Recall	**0.974**	**0.857**	**0.963**	0.571	**0.979**	**0.381**
	F-measure	**0.971**	**0.867**	**0.936**	**0.658**	**0.925**	**0.516**
	Accuracy	**0.952**		**0.892**		**0.87**	

Conclusion

Optimizing local predictors by propagating beliefs over a Markov Random Field permits to correct misclassified labels. This improves the performance of baseline supervised classifiers even when these classifiers are weak.

Classifiers (SVM and MRF) trained using the set of features we specifically selected for the task of social spammers detection on Twitter, have a significantly better performance than those trained using the sets of Benevenuto's and Stringhini's.

As a future work, we would like to investigate the feasibility of using our findings to design adaptive classifiers based on graphical inference.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, p. 12, 2010.

[2] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 1–9, ACM, 2010.

[3] N. El-Mawass, P. Honeine, and L. Vercouter, "Supervised classification of social spammers using a similarity-based markov random field approach," in *Proceedings of the 5th Multidisciplinary International Social Networks Conference (MISNC 2018)*, ACM, 2018.

[4] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," *arXiv preprint arXiv:1701.03017*, 2017.