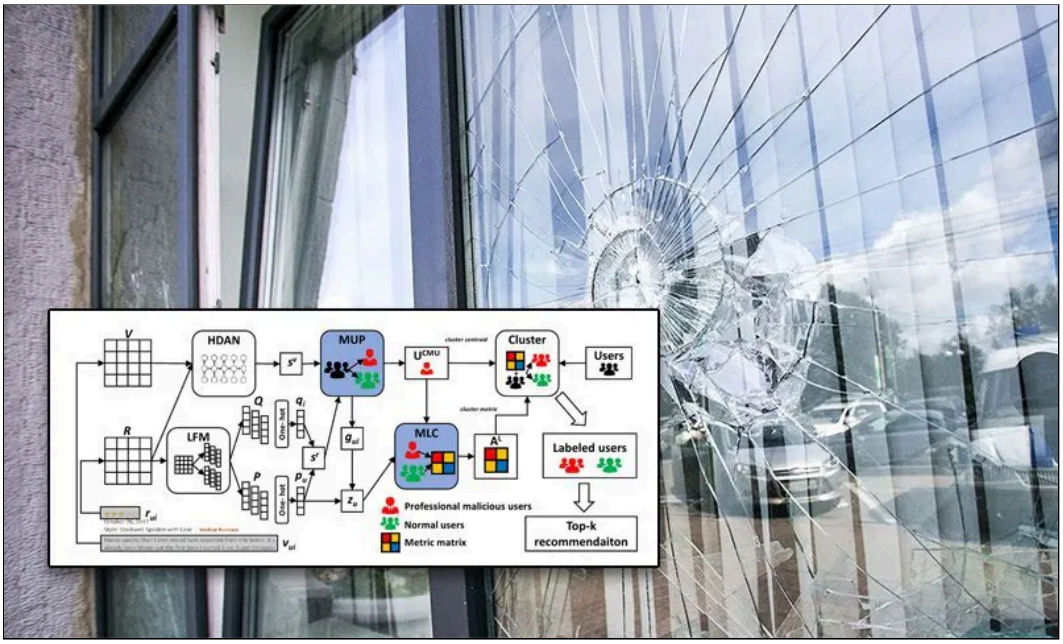
Detecting ‘Professional’ Malicious Online Reviews with Machine Learning

Originally published at https://www.unite.ai/detecting-professional-malicious-online-reviews-with-machine-learning/



A new research collaboration between China and the US offers a way of detecting malicious ecommerce reviews designed to undermine competitors or to facilitate blackmail, by leveraging the signature behavior of such reviewers.

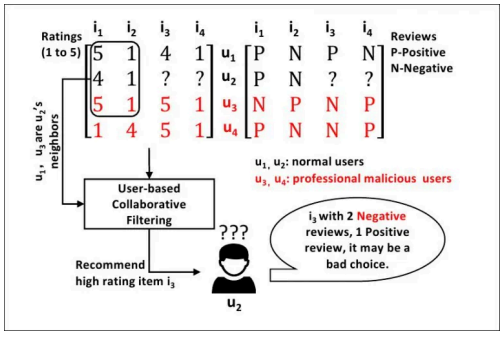
The system, titled *malicious user detection model* (MMD), utilizes Metric Learning, a technique commonly used in computer vision and recommender systems, together with a Recurrent Neural Network (RNN), to identify and label the output of such reviewers, which the paper names *Professional Malicious Users* (PMUs).

Great! 1 star

Most online ecommerce reviews provide two forms of user feedback: a star rating (or a rating out of 10) and a text-based review, and in a typical case, these will correspond logically (i.e., a bad review will be accompanied by a low rating).

PMUs, however, typically subvert this logic, by either leaving a bad text review with a high rating, or a poor rating accompanied by a good review.

This allows the user’s review to cause reputational damage without triggering the relatively simple filters deployed by ecommerce sites to identify and address the output of maliciously negative reviewers. If a filter based on Natural Language Processing (NLP) identifies invective in the text of a review, this ‘flag’ is effectively cancelled by the high star (or decimal) rating that the PMU also assigned, effectively rendering the malicious content ‘neutral’, from a statistical point of view.



An example of how a malicious review can be commingled, statistically, with genuine reviews, from the point of view of a collaborative filtering system that’s trying to identify such behavior.
Source: https://arxiv.org/pdf/2205.09673.pdf

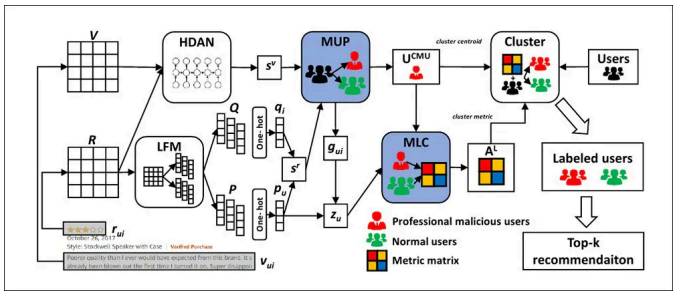
The new paper notes that the intention of a PMU is often to extort money from online retailers in return for amendment of negative reviews, and/or a promise to post no further negative reviews. In some cases, the actors are *ad hoc* individuals seeking discounts, though frequently the PMU is being casually employed by the victim's competitors.

Cloaking Negative Reviews

The current generation of automated detectors for such reviews use Collaborative Filtering or a content-based model, and are looking for clear and unambiguous 'outliers' – reviews which are uniformly negative across both feedback methods, and which diverge notably from the general trend of review sentiment and rating.

The other classic signature that such filters key on is a high posting frequency, whereas a PMU will post strategically and only occasionally (since each review may represent either an individual commission, or a stage in a longer strategy designed to obfuscate the 'frequency' metric).

Therefore the new paper's researchers have integrated the strange polarity of professional malicious reviews into a dedicated system, resulting in an algorithm that's almost on a par with the ability of a human reviewer to 'smell a rat' at the disparity between the rating and the review text content.



The conceptual architecture for MMD, comprised of two central modules: Malicious User Profiling (MUP) and Attention Metric Learning (MLC, in grey).

Comparison to Prior Approaches

Since MMD is, the authors state, the first system to attempt to identify PMUs based on their schizophrenic posting style, there are no direct prior works against which to compare it. Therefore the researchers pitted their system against a number of component algorithms on which traditional automated filters frequently depend, including K-means++ Clustering; the venerable Statistic Outlier Detection (SOD); Hysad; Semi-sad; CNN-sad; and Slanderous user Detection Recommender System (SDRS).

	Amazon			Yelp		
	SEN	*SPE*	F-score	*SEN*	*SPE*	F-score
K-means++	0.381	0.706	0.494	0.201	0.773	0.318
SOD	0.062	0.984	0.113	0.041	0.962	0.076
Hy-sad	0.371	0.853	0.516	0.542	0.889	0.671
Semi-sad	0.442	0.784	0.563	0.661	0.933	0.773
CNN-sad	0.552	0.791	0.650	0.663	0.922	0.771
SDRS	0.651	0.874	0.745	0.764	0.913	0.831
MLC	0.861	0.967	0.910	0.821	0.938	0.874
MUP	0.662	1*	0.795	0.742	1*	0.850
MMD (MUP+MLC)	**0.921***	**0.970**	**0.944***	**0.98***	**0.996**	**0.987***

Tested against labeled datasets from Amazon [securities_stock_price_tag symbol="amzn" exchange="nasdaq"] and Yelp, MMD is able to identify professional online detractors with the highest rate of accuracy, the authors claim. Bold represents MMD, while the asterisk (*) indicates the best performance. In the above case, MMD was beaten in only two tasks, by a standalone technology (MUP) that is already incorporated into it, but which is not tooled by default for the task at hand.

	Taobao			Jingdong		
	SEN	*SPE*	F-score	*SEN*	*SPE*	F-score
K-means++	0.141	0.728	0.234	0.44	0.667	0.530
SOD	0.022	0.978	0.039	0.14	0.896	0.242
Hy-sad	-	-	-	-	-	-
Semi-sad	-	-	-	-	-	-
CNN-sad	-	-	-	-	-	-
SDRS	0.451	0.884	0.597	0.732	0.911	0.811
MLC	-	-	-	-	-	-
MUP	0.641	0.996*	0.779	0.62	0.998*	0.764
MMD (MUP+MLC)	**0.941***	**0.987**	**0.963***	**0.96**	**0.989**	**0.974***

In this case, MMD was pitted against unlabeled datasets from Taobao and Jindong, making it effectively an unsupervised learning task. Again, MMD is only improved upon by one of its own constituent technologies, highly adapted for the task for the purpose of testing.

The researchers observe:

> *'[On] all four datasets, our proposed model MMD (MLC+MUP) outperforms all the baselines in terms of F-score. Note that MMD is a combination of MLC and MUP, which ensures its superiority over supervised and unsupervised models in general.'*

The paper also suggests that MMD could serve as a useful pre-processing method for traditional automated filter systems, and provides experimental results on a number of datasets, including User-based collaborative Filtering (UBCF), Item-based collaborative Filtering (IBCF), Matrix Factorization (MF-eALS), Bayesian personalized ranking (MF-BPR), and Neural Collaborative Filtering (NCF).

In terms of Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) in the results of these tested augmentations, the authors state:

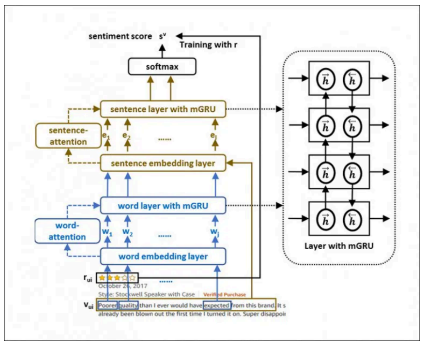
> *'Among all four datasets, MMD improves the recommendation models significantly in terms of HR and NDCG. Specifically, MMD can enhance the performance of HR by 28.7% on average and HDCG by 17.3% on average.*
>
> *'By deleting professional malicious users, MMD can improve the quality of datasets. Without these professional malicious users' fake* [feedback]*, the dataset becomes more* [intuitive]*.'*

The paper is titled *Detect Professional Malicious User with Metric Learning in Recommender Systems*, and comes from researchers at the Department of Computer Science and Technology at Jilin University; the Key Lab of Intelligent Information Processing of Chinese Academy of Science at Beijing; and the School of Business at Rutgers in New Jersey.

Data and Approach

Detecting PMUs is a multimodal challenge, since two non-equivalent parameters (a numerical-value star/decimal rating and a text-based review) must be considered. The authors of the new paper assert that no prior work has addressed this challenge.

MMD employs a Hierarchical Dual-Attention recurrent Neural network (HDAN) to assimilate the review content into a sentiment score.



Projecting a review into a sentiment score with HDAN, which contributes word embedding and sentence embedding in order to obtain a sentiment score.

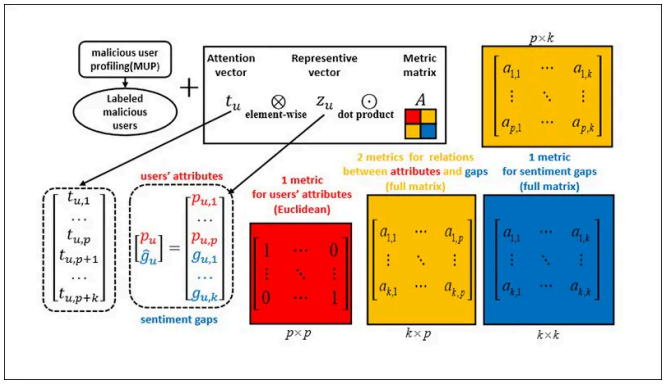
HDAN uses attention mechanisms to assign weights to each word, and to each sentence. In the image above, the authors state, the word *poorer* should clearly be assigned greater weight than competing words in the review.

For the project, HDAN took the ratings for products across four datasets as ground truth. The datasets were Amazon.com; Yelp for RecSys (2013); and two 'real world' (rather than experimental) datasets, from Taobao and Jindong.

MMD leverages Metric Learning, which attempts to estimate an accurate distance between entities in order to characterize the overall group of relationships in the data.

MMD begins with a one-hot encoding to select the user and item, via a Latent Factor Model (LFM), which obtains a base rating score. In the meantime, HDAN projects the review content into the sentiment score as adjunct data.

The results are then processed into a Malicious User Profiling (MUP) model, which outputs the *sentiment gap vector* – the disparity between the rating and the estimated sentiment score of the review's text content. In this way, for the first time, PMUs can be categorized and labeled.



Attention-based Metric Learning for clustering.

Metric Learning for Clustering (MLC) uses these output labels to establish a metric against which the probability of a user review being malicious is calculated.

Human Tests

In addition to the quantitative results detailed above, the researchers conducted a user study that tasked 20 students with identifying malicious reviews, based only on the content and star rating. The participants were asked to rate the reviews as *0* (for 'normal' reviewers) or *1* (for a professional malicious user).

Out of a 50/50 split between normal and malicious reviews, the students labeled 24 true positives and 24 true negative users on average. By comparison, MMD was able to label 23 true positive and 24 true negative users on average, operating almost at human-level discernment, and surpassing the baselines for the task.

	Taobao		
	SEN	*SPE*	F-score
K-means++	0.52	0.6	0.557
Semi-sad	0.68	0.76	0.718
MLC	0.92	0.88	0.899
MUP	0.76	1*	0.867
MMD (MUP+MLC)	**0.92**	**0.96**	**0.94**
Students	0.96*	0.96	0.96*

Students vs. MMD. Asterisk [*] indicates best results, and bold indicates MMD's results.

The authors conclude:

> *'In essence, MMD is a generic solution, which can not only detect the professional malicious users that are explored in this paper but also serve as a general foundation for malicious user detections. With more data, such as image, video, or sound, the idea of MMD can be instructive to detect the sentiment gap between their title and content, which has a bright future to counter different masking strategies in different applications.'*

First published 20th May 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai