Отчёт Завадского Кости
Задача №2, этап L2
Поиск фейковых отзывов

1 Поиск научных статей

Среди всех возможных научных статей рассматривались статьи, у которых в abstract и названии присутствовало хотя бы одно из следующих 4-ёх слов:

- "review",
- "product",
- "fake",
- "amazon",
- "spam",
- "detection".

На основе abstract и названия были отобраны следующие релевантные статьи:

1. Discussion about Attacks and Defenses for Fair and Robust Recommendation System Design;
2. Explainable Verbal Deception Detection using Transformers;
3. Detecting Spam Reviews on Vietnamese E-commerce Websites;
4. Opinion Spam Detection: A New Approach Using Machine Learning and Network-Based Algorithms;
5. Fake or Genuine? Contextualised Text Representation for Fake Review Detection;
6. Social Fraud Detection Review: Methods, Challenges and Analysis;
7. Confounds and Overestimations in Fake Review Detection: Experimentally Controlling for Product-Ownership and Data-Origin;
8. Identifying Hijacked Reviews;
9. Fake Reviews Detection through Ensemble Learning;
10. Robust Spammer Detection by Nash Reinforcement Learning;
11. Fake Review Detection Using Behavioral and Contextual Features;
12. ColluEagle: Collusive review spammer detection using Markov random fields;
13. Opinion Spam Recognition Method for Online Reviews using Ontological Features;
14. Opinion Fraud Detection via Neural Autoencoder Decision Forest;
15. Credible Review Detection with Limited Information using Consistency Analysis;
16. BIRDNEST: Bayesian Inference for Ratings-Fraud Detection;
17. Amazon Fake Reviews;
18. Spatio-Temporal Graph Representation Learning for Fraudster Group Detection;
19. Data Poisoning Attacks to Deep Learning Based Recommender Systems;
20. Amazon Fake Reviews;
21. Wide-Ranging Review Manipulation Attacks: Model, Empirical Study, and Countermeasures;

22. What Yelp Fake Review Filter Might Be Doing;
23. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors;
24. Fraudulent user prediction in rating platforms;
25. Spotting opinion spammers using behavioral footprints;
26. Review Spam Detection via Temporal Pattern Discovery;
27. Bounding Graph Fraud in the Face of Camouflage;
28. Combating crowdsourced review manipulators: A neighborhood-based approach;
29. Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection;
30. Collective Opinion Spam Detection: Bridging Review Networks and Metadata.

Указанные статьи были распределены по важности на основе уровня конференции, на которых статьи докладывались. В следующем списке указаны уровни конференций, на которых докладывались статьи и дата публикации:

Рейтинг конференций, на которых докладывались статьи			
Номер и название статьи	Название конференции	Ранг	Дата
3. Detecting Spam Reviews on Vietnamese E-commerce Websites	Asian Conference on Intelligent Information and Database Systems	*B*	9 Dec 2022
8. Identifying Hijacked Reviews	International Joint Conference on Natural Language Processing	*B*	7 Jul 2021
10. Robust Spammer Detection by Nash Reinforcement Learning	KNOWLEDGE DISCOVERY AND DATA MINING	A^*	22 Jun 2020
19. Data Poisoning Attacks to Deep Learning Based Recommender Systems	Usenix Network and Distributed System Security Symposium	A^*	8 Jan 2021
21. Wide-Ranging Review Manipulation Attacks: Model, Empirical Study, and Countermeasures	CIKM	*A*	2019
23. Handling cold-start problem in review spam detection by jointly embedding texts and behaviors	ACL	A^*	2017
24. Fraudulent user prediction in rating platforms	WSDM	A^*	2018
25. Spotting opinion spammers using behavioral footprints	KDD	A^*	2013
26. Review Spam Detection via Temporal Pattern Discovery	KDD	A^*	2012
27. Bounding Graph Fraud in the Face of Camouflage	KDD	A^*	2016
28. Combating crowdsourced review manipulators: A neighborhood-based approach	WSDM	A^*	2018
29. Alleviating the Inconsistency Problem of Applying Graph Neural Network to Fraud Detection	SIGIR	A^*	2020
30. Collective Opinion Spam Detection: Bridging Review Networks and Metadata	KDD	A^*	2015

все остальные статьи неопубликованы.

2 Обзор статей

В ходе изучения указанного списка статей были выявлены сильные недостатки многих статей. В основном они заключались в

- предположении, что различные спамеры пользуются одним и тем же алгоритмом;
- том, что спамеры не читают научные статьи (то есть метод подвержен атакам со стороны спамеров);
- том, что в статье не ставилась бизнес задача (оптимизировались метрики AUC, top-k),
- том, что статья 3 использовала данные вьетнамского магазина (покупатели вьетнама намного беднее покупателей России, а значит в этой статье изучается совсем другое явление).

На основе выше сказанного было принято решение изучить подробно статьи 8, 10, 19 и 29. Именно на основе этих статей и будет основано решение.

3 Robust Spammer Detection by Nash Reinforcement Learning

Есть граф, состоящий из узлов – продуктов (V) и аккаунтов (U), а также рёбер (появляется, если аккаунт u_i оставляет отзыв на продукт v_j). Ребру приписывается набор характеристик $x_{i,j}$(продукта, аккаунта, времени и т.д.). Спамеры используют различные алгоритмы спамов A (в какое-то временное окно решают какие рёбра графа создавать и с какими характеристиками (в расположении спамера есть новые аккаунты и элитные)). Mixed attack strategy – это

$$A(p) = \mathbb{E}_{k\sim p}(a_p) = \sum p_k a_k,$$

где P – распределение вероятностей на A.

Из-за разнообразия спам стратегий используют различные детекторы $[d_1, \ldots, d_L]$. Для каждого детектора определяют важность $[q_1, \ldots, q_L]$. Detector in effect это

$$D(q) = \sum_{i=1}^{L} d_i q_i.$$

Будем предполагать, что задана самая сильная стратегия спама, которую спамеры могут разработать (учитывая знания из некоторых статей). Это предположение имеет больше практического смысла: профессиональные спамеры могут получить доступ к деталям детекторов спама через опубликованные статьи, обратную разработку из помеченных спамов и объяснения обнаружения, и обновить свои стратегии рассылки спама, чтобы обойти фиксированный детектор. К тому же дефолтный детектор можно обмануть спамом в камуфляже.

Задача статьи выглядит так

$$\min_q \max_p \sum_{v\in V_T} \max\{0, PE(v, R, p, q)\},$$

где Practical Effect (PE) of spamming using $A(p)$ on v against the detection of $D(q)$. Решение этой задачи приведёт к созданию детектора, способного противостоять любым смешивающимся спамерским стратегиям с весами $\alpha = [\alpha_1, \ldots, \alpha_K]$. В частности мы стремимся получить универсальный детектор, который будет минимизировать практические эффекты спама.

Функция PE не дифференцируема! Такая функция PE надёжно обнаружает спам. В качестве обучения будет использоваться стахостическую оптимизацию на основе Монте-Карло для решения задачи. Два игрока будут играть до тех пор, пока не достигнут равновесия Нэша. На практике равновесие Нэша достижимо.