Effective Listing Spam Detection System using Locality Sensitive Hashing at Scale

Chandan Maloo, Akhil Kaza
[*] Offerup

Abstract

The popularity, cost-effectiveness and ease of buying and selling that marketplaces like Craigslist, Offerup offer to users has been plagued with the rising number of unsolicited spam listings, fraudulent transactions and in some extreme cases law enforcement also needs to be involved. Driven by the need to protect Offerup users from this growing menace, research in spam, fraud listing filtering/detection systems has been increasingly active in the last decade. However, the adaptive nature of Scammers and Fraudsters has often rendered most of these systems ineffective. While several spam detection models have been reported in literature, the reported performance on an out of sample test data shows the room for more improvement. Presented in this research is an improved spam detection model based on Locality Sensitive Hashing algorithm which to the best of our knowledge has received little attention in spam/fraud detection problems. Experimental results show that the proposed model outperforms earlier approaches across a wide range of evaluation metrics inside Offerup.

Index Terms- Marketplace Spam, Marketplace Fraud, Machine Learning, Spam Detection, Human Intelligence, LSH, Locality Sensitive Hashing

1. *Introduction*

Marketplaces like Offerup, Craigslist excel at providing customers ability to sell/buy unused goods from the comfort of your home. There is a saying "one man's trash is another man's treasure" and these marketplaces provide an opportunity to make this happen. Customers post daily millions of listings on these platforms selling products, services and various other activities helping folks earn livelihood and well as get what they look for, helping our planet be more sustainable by promoting reuse, recycling. Despite these positives, these platforms are generally plagued with unsolicited and occasionally fraudulent listings popularly called spam or junk or frauds listings. A reasonable portion of them are disguised to maliciously mislead or sometimes defraud the recipient; threatening the essence of these marketplaces as reliable and trustworthy to shop. The continued rise in spam listings has inspired mitigative approaches to protect Marketplaces users by filtering or rejecting them all together. Thus, the need for spam filtering which basically entails isolating spam from non-spam listings using computational tools. An even more important stage in the filtering process is detecting that a listing is spam or not because this determines what is done to the listing afterwards. Accurate detection has been an active area of research with marked rise in the use of computational intelligent methods in the last decade [2]. However, spammers have increasingly over the years adapted spam listings to appear like legitimate ones. Consequently, the adaptive nature of spam listings have often deceived some of the most effective spam filters; hence the continued need for more accurate spam detection tools still remains an open area of research [1]. Although many spam detectors have been reported in literature, the predictive accuracy in the various researches suggests the need for better methods with improved accuracy. Bayesian classifiers are one of the earliest methods for spam email filtering and recommendations have been made to facilitate its viability for practical use [4]. For instance, In a more recent study, [3] carried out a comparative analysis of 14 different machine learning algorithms and Rotation Tree algorithms performed best on an out of sample test set. Other algorithms like J48, Bayesian Logistic Regression and Multi-Layer Perceptron were also reported with relatively good performance. Meanwhile, [1] introduced a Support Vectors Machine (SVM)-based spam detector using the same dataset under similar experimental setting and obtained a better accuracy on the testing set compared to previous works on the same dataset. Locality Sensitive Hashing[5] is an efficient

and scalable algorithm that has found widespread application in several Machine learning competitions. We summaries the contribution of this research as follows

- We propose an improved listing spam detection system using Locality Sensitive Hashing.
- We compare the performance of the proposed system with previous works on the same data

  We analyze the performance of the proposed system using a wider range of evaluation metrics beyond accuracy which has dominated the previous studies

The remaining part of this manuscript is structured as follows. Section 2.0 describes the materials and methods of this research which features a description of the proposed Locality Sensitive Hashing based model, the dataset, evaluation metrics and the experimental design and model implementation details. The experimental results are discussed in section 2.2. while the research is concluded in section 3.0.

2. **2. *Materials and Methods***

Offerup collects a lot of data from its customers who are interested in listing their products or would like to purchase a product from the marketplace. We utilize a range of information to build our service, the core signals are User Profiles and User Reports. User Profile captures aggregate quantitative factors like # of listings, sell rate, time on Offerup, ratings, reviews and 100 other keys stats. User Reports is like an Andon Cord[1] mechanism provided to Offerup Customers to report about the User or the listings. We use a combination of these two to create signatures for our customers. When a listing is reported as Spam by our customers we start collecting this information in a datastore. Another system we leverage is a Human Review system similar to Squad[2]/ Mechanical Turk which provides us confidence that the spam reported is indeed spam. Armed with this data we have collected millions of listings which have been marked spam/fraud by the customers.
This data is now collected on a regular basis and passed through our LSH[3] systems to generate signatures. We maintain these signatures a big cache for faster retrieval. When an input listing comes we search in our Signatures cache and if it matches >=3 signatures with 70% match we mark the listing as SPAM to be reviewed. We have various levels of rules to mark them auto SPAM or pass for Human Judgement.

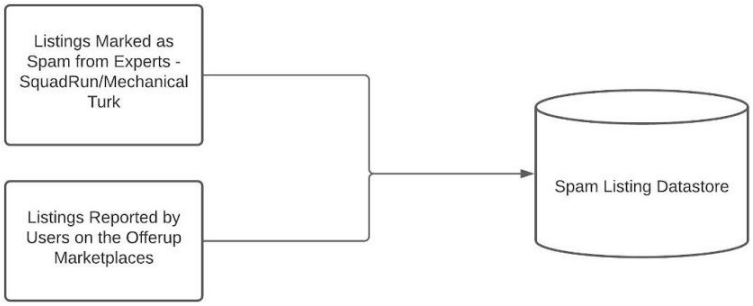
2.1: How do we generate LSH signatures

We maintain two types of signatures Listing Signature and User Signature. Listed below is a methodology on how we generate listing signatures.
Every listing has the following attributes, user, listing title, description, price, location, images. We extract each of these information and make a Signature store for all them. In production we have a Listing Title, Description Store and we are working on Image Store. We basically take ngrams of our title and description and then use techniques like TFIDF, Shingling all part of the LSH process to generate signatures.

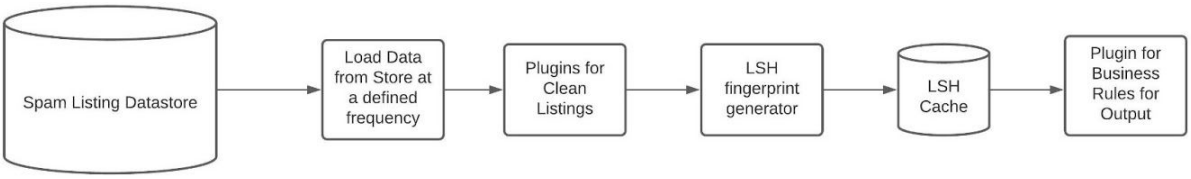
The final system is modular and is composed of 3 stages 1) Data Generation 2) Offline Processing Stage 3) Inference Stage for actual real time inference

[1] https://en.wikipedia.org/wiki/Andon_(manufacturing)
[2] https://www.squadstack.com/
[3] https://en.wikipedia.org/wiki/Locality-sensitive_hashing

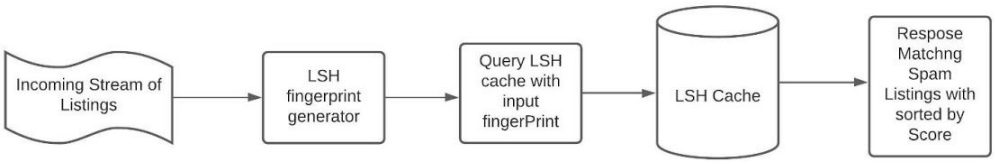
1. Data Generation Stage



2. Processing Stage



3. Inference Stage



2.2. *Results*

Below runs shows that when learning data on a given day the LSH model is able to capture 50% of the SPAM listings based on them being seen earlier. While in the good population we do not have any false positives.

Sample Data Date	***Good Examples***	***Detection***	***%Detection***	***Bad/Spam Examples***	***Detection***	***%Detection***
Week ending 2020-05-11	*189*	*5*	*2.65%*	*Model learned from this set*		

Week ending 2020-05-19	*159*	*4*	*2.52%*	*323*	*202*	62.54%
Week ending 2020-05-19 sample2				*298*	*165*	55.37%
Week ending 2020-05-06	*179*	*1*	*0.56%*	*504*	*290*	57.54%
Week ending 2020-05-06				*540*	*300*	55.56%

3. *Conclusion*

As one can observe, LSH cache along with Human intelligence can help us set up an adaptive Spam Listing system that can mitigate spam listings and help maintain our marketplaces. Using n-grams we are able to take advantage of spelling variations that Scamsters/Spammers use to trick folks to doing the intended bad activity. By sharing the SPAM result with the user we are able to collect data on the efficiency of the whole system. We also present a modular architecture that can house multi-tenant datastores and scale the system to millions of listings. Making sure we keep refreshing our datasets guarantee that we are able to take known Scam from our systems and provide users with a good customer experience.

4. *References*

1. Olatunji, S.O., Improved email spam detection model based on support vector machines. Neural Computing and Applications, 2017: p. 1-9.
2. Bhuiyan, H., et al., A Survey of Existing E-mail Spam Filtering Methods Considering Machine Learning Techniques. Global Journal of Computer Science and Technology, 2018.
3. Shuaib, M., et al., Comparative Analysis of Classification Algorithms for Email Spam Detection.International Journal of Computer Network and Information Security, 2018. 10(1): p. 60.
4. Androutsopoulos, I., et al., An evaluation of naive bayesian anti-spam filtering. arXiv preprint cs/0006013, 2000.
5. Online Generation of Locality Sensitive Hash Signatures https://www.aclweb.org/anthology/P10-2043.pdf