JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Anomaly Detection of Products in e-Commerce Exchange using Artificial Intelligence for Demand Insights

K. Victor Rajan, Freddy Frejus
Department of Computer Engineering, Atlantic International University
Honolulu, Hawaii 96813, USA

Abstract: Online e-commerce exchanges enable the exchange of products and services across organizational or individual boundaries through internet. Transactions are digitally enabled among organizations and individuals. E-commerce exchanges execute a very large number of sales updates when compared to brick-and-mortar stores. Even a few misinterpreted items can have a significant business impact and result in wrong demand/marketing insights. Multiple vendors might sell similar products with slightly different product name or description. Early detection of anomalies in an automated real-time fashion is an important need for such exchanges to predict demand with higher accuracy. Manufacturers and suppliers depend on exchange data to estimate demand for a product. Generating demand insight for a particular type of product is a challenging task since vendors do not use common nomenclature for product names or unique product codes globally. Human error also leads to the sale of same product under anomalous names. Many e-commerce exchanges use domain experts to review the product details and group them together before business intelligence reports are generated. With the use of internet growing, e-commerce exchanges see millions of products flowing in every day. Employing man power for matching of products is time consuming and costly as the volume is huge. In this paper, we describe AI-based anomaly detection approach we developed and evaluated for a large-scale online exchange. Our system detects anomalies both in batch and real-time streaming settings, and the items flagged for manual review are found to be very less after automatic processing.

Index Terms **- E-commerce Exchange, Anomaly Detection, Demand Insights.**

I. Introduction

Over the last decade, the advent of e-commerce has actually transformed the manner in which people buy products. People now are not only just using internet for gathering information, leisure or socializing online but also seek measures to conduct business online. Even popular social networking sites like Face book are allowing people to promote and sell products and services online. Introduction of computer and mobile based e-commerce application software provides evidence of how e-commerce has boomed over the past 5 years. This provides us enough amounts of data to generate demand insights for future planning. However, there is high level of human error introduced while labeling the final product due to anomalies in product name and description. Hardly a manufacturer wants to create an anomalous product. But, manufacturer might change the name slightly in due course to make it attractive or give useful meaning. Many manufacturers might produce the same product but sell under slightly different name. Hence, the name anomalies are like a needle in a haystack which renders significantly imbalanced dataset. An e-commerce exchange which uses the historical data for business intelligence might get estimates for same product under multiple item names. Grouping and consolidation are done manually by experts before this data is sent for analytics. Automating this matching activity will save time and cost and make way for real time analysis. Currently, ambiguous items are manually processed, reviewed, merged with main stream data and then sent for analytics. In this research paper, we first analyze the problems, huddles arising due to product anomalies. Then we formally define the problem and identify a use case for experiment. We used the medical products from a health care giant. We used data science and machine learning to cleanse the data, remove anomalies and group the products together based on product names. This gives better business insights for analysts about sale of specific product from multiple vendors in an exchange. The remainder of this paper is organized as follows. In section 2, we formally formulate the problem and pick up a case study for analysis. We propose an AI-based methodology for anomaly detection in Section 3. In Section 4, the results of our experiments are described. Finally, the paper is concluded in Section 5.

II. Problem Definition And Related Work

Anomaly detection is the identification of items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically, anomalous data can be connected to some kind of problem or rare event or ambiguous products. This connection makes it very interesting to be able to pick out which data points can be considered anomalies, as identifying these events are typically very interesting from a business perspective. This is why many e-commerce and data-driven companies have adopted automated anomaly detection. In e-commerce, nothing is ever static; products are constantly added and removed; discounts and rebates from manufacturers and retailers change, competitors attempt to out-maneuver each other with ad buys and social media marketing campaigns, and so on. With all these moving pieces, the last thing your business needs is to be overwhelmed by alert storms or burdened with setting and then constantly adjusting static thresholds, but traditional business intelligence (BI) tools that rely on traditional anomaly detection leave you no other choice than setting static thresholds or tracking the dashboards yourself. We used data from a popular healthcare exchange for our analysis and experiment. The exchange allows suppliers to list their products and buyers make fixed term agreement for bulk procurement. Their purchase order information is fed to the data analytics software to generate business intelligence dashboards. The product anomaly is a big challenge to analyze and predict demand for any specific product. Following table contains examples of few anomalous products we picked from health care exchange.

No	Product Description	Unit of Measurement	Conversion Factor
1	WIRE, BIOSCREW HYPERFLEX GUIDE .045	BX	5
2	WIRE, BIOSCREW HYPERFLEX GUIDE .045	CS	5
3	MARKER,SKIN	EA	10
4	MARKERS,SKIN & RULER,PR INK	CA	100
5	MASK,RESP & SURG,VFLEX,N95,SM	EA	1
6	MASK,RESPIRATOR STD N95	KT1	1
7	1.3MM CORTEX SCREW SLF-TPNG WI	EA	1
8	1.3MM CORTEX SCREWS SELF TAPPI	EA	1
9	1.3MM TI CRTX SCR SLF-TPNG W/P	EA	1
10	1.3MM SELF TAPPING CORTEX SCREW 18MM	EA	1

Table 1: Health Care Products with Anomalies

We observed that anomalies are introduced due to the following reasons:

- Manufacturers use different notations for Unit of Measurement (UOM).
- Use of abbreviations, acronyms, and special characters in names, descriptions.

The analytics dashboard generated based on the above details would result in incorrect insights. Items 1 and 2 will be classified as different products, but they are of same type. Items 5 and 6 will also be distinguished as different products. Items 9 and 10 will be classified as different, but all four items from 7-10 are same. Sales, revenue forecasts made from this data warehouse will mislead the decision makers unless such anomalies are removed. Organizations use experts to remove anomalies before sending the data for analytics. Health care experts review the product details and group them together by assigning labels or tags. However this involves huge man power cost and batch processing. Quick real time decision cannot be taken.

Research on anomaly detection in e-commerce and retail business is growing nowadays due to demand for automated analytics. Unfortunately, existing anomaly detection algorithms usually focus on single attribute like price, category etc. Multi-attribute anomaly detection for products is still evolving. Anomaly detection for an e-commerce pricing system was presented by Shaabani et al [1]. While there is some work on anomaly detection in retail [7], to the best of our knowledge, there does not seem to be references on anomaly detection for a large-scale multi-attribute system as we have considered in this paper. Much of the literature also focuses on time-series anomaly detection approaches. The use of anomaly detection methods in production systems is prevalent among large technology companies like Google, Yahoo but not in e-commerce companies. We mainly focus on contextual anomalies because we are interested in whether an item is an anomaly at a particular point in time.

III. Research Methodology

Anomalous products in e-commerce exchange make the warehouse data non-suitable for forecasts and predictions. We want to identify and correct data errors that are the root cause of a product anomaly. This includes item attributes such as descriptions and unit of measurement (UOM) from purchase invoices. There is no unique product ID for health care products. Products are normally identified using description, UOM, and conversion factor. Our anomaly detection system first produces a feature vector by combining the three attributes.

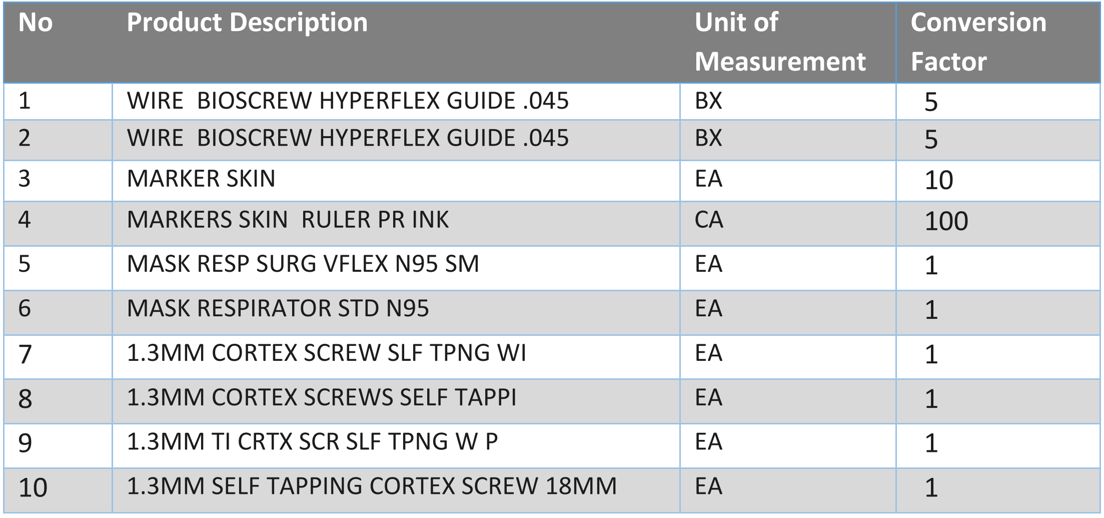
The Feature Extractor (FE) works as follows.

- Identify UOM mismatch using a table look up. Replace vendor UOM with exchange UOM using the mapping as shown in Table 2.
- Special characters like dollars, brackets, ampersands, commas etc., are removed from product description.
- All words are converted to uppercase.

EXCHANGE UOM	EQUIVALENT VENDOR UOM
EA	KT1, UT, SN, IN
CA	CN, CRN, CT
BX	CS, CAS, BOX

Table 2: UOM translation

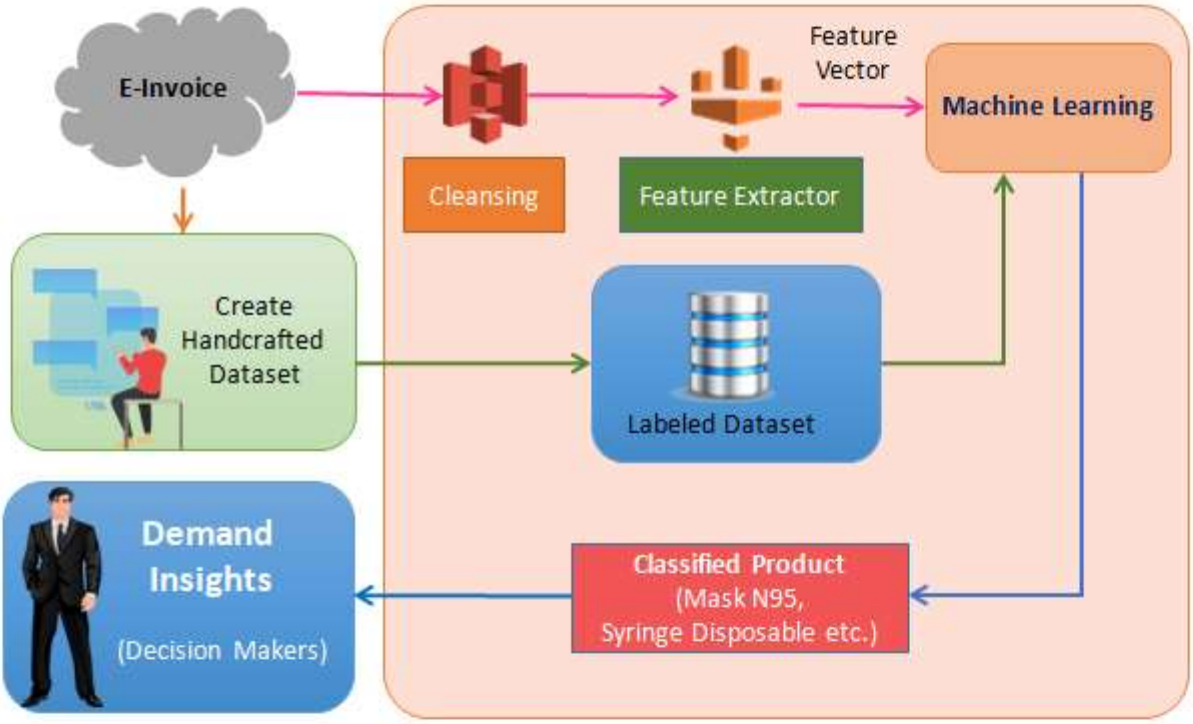
The translated attributes for the anomalous products in Table 1 are shown below for easy understanding.



No	Product Description	Unit of Measurement	Conversion Factor
1	WIRE BIOSCREW HYPERFLEX GUIDE .045	BX	5
2	WIRE BIOSCREW HYPERFLEX GUIDE .045	BX	5
3	MARKER SKIN	EA	10
4	MARKERS SKIN RULER PR INK	CA	100
5	MASK RESP SURG VFLEX N95 SM	EA	1
6	MASK RESPIRATOR STD N95	EA	1
7	1.3MM CORTEX SCREW SLF TPNG WI	EA	1
8	1.3MM CORTEX SCREWS SELF TAPPI	EA	1
9	1.3MM TI CRTX SCR SLF TPNG W P	EA	1
10	1.3MM SELF TAPPING CORTEX SCREW 18MM	EA	1

Table 3: Product attributes after feature extraction

The revised attributes are passed to machine learning algorithm to remove anomalies and match the products. The following diagram shows the architecture of our proposed system.



. Figure 1: Architecture of Anomaly Detection System

Pairs of feature vectors and product names are fed into the machine learning algorithm to generate a model. The well-trained model produces product suggestions (Mask N95, Syringe Disposable, etc.) for a given input.

3.1Population and Sample

In deep learning, supervised deep anomaly detection involves training a deep supervised binary or multi-class classifier, using labels of both normal and anomalous data instances.

Supervised deep anomaly detection models, formulated as multi-class classifier aids in detecting rare brands, new drug name etc. We consider the problem of anomaly detection with a small set of partially labeled anomaly examples and a large-scale unlabeled dataset. This is achieved by jointly optimizing the exploitation of the small labeled training data and the exploration of the rare unlabeled anomalies. We use neural network with L hidden layers and corresponding set of weights W = {W1, W2, … WL}. The objective is to train the neural network to learn a transformation that minimizes the volume of a data-enclosing hyper sphere in output space Z centered on a predetermined point C. Penalizing the mean squared distance of the mapped samples to the hyper-sphere center C forces the network to extract those common factors of variation which are most stable within the dataset. As a consequence normal data points tend to get mapped near the hyper-sphere center, whereas anomalies are mapped further away. After initialization, the hyper-sphere center C is set as the mean of the network outputs obtained from an initial forward pass of the data. Once the network is trained, the anomaly score for a test point x is given by the distance from $\phi(\boldsymbol{x}; \mathcal{W})$ to the center of the hyper-sphere:

$$s(\boldsymbol{x}) = \|\phi(\boldsymbol{x}; \mathcal{W}) - \mathbf{c}\|$$

Sentence vectors are converted to numerical vectors for calculating distance between them. Word embedding techniques are used to represent words mathematically. TF-IDF, Word2Vec, FastText are frequently used Word Embedding methods. We use Term Frequency-Inverse Document Frequency (TF-IDF) method to find the similarity between two documents. TF-IDF method determines the relative frequency of words in a specific document through an inverse proportion of the word over the entire document corpus. In TF-IDF, similar text must result in closer vector. TF-IDF is the product of the TF and IDF scores of the term.

TF = number of times the term appears in the doc/total number of words in the document.

$$f_{ij} = \textit{frequency of term } i \textit{ in document } j$$

IDF = ln (number of docs/number docs the term appears in)

$$idf_i = log_2\left(\frac{N}{df_i}\right)$$

Hence,

$$\text{TF-IDF} = tf_{ij}idf_i = tf_{ij} \times log_2\left(\frac{N}{df_i}\right)$$

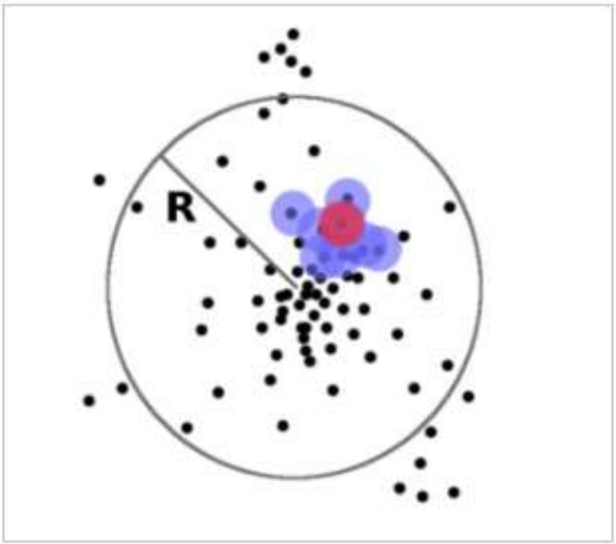


Figure 2: Anomaly Detection using Cluster Strategy

Our strategy queries data points that are likely anomalies in clusters near the decision boundary. The anomaly radius R is used to decide the boundary for the cluster. Firstly, we obtained the prior anomaly score of labeled data via prior estimation module and attach the prior anomaly score to unlabeled data as the training weight. The anomaly scoring neural network uses the optimal value R to ensure that significantly deviated data items are not clustered into the hyper-sphere. Items having distance greater than R from the centre of the cluster are excluded from the hyper-sphere.

We discuss the experiment and results in the following section.

IV. Experiment And Results

We deployed our model in a reputed health care exchange and tested. We had a daily job that applied our anomaly detection system on the entire invoice catalog. If items were anomalous and had a high business impact, they were sent for review. The alerted anomalies can be viewed through a web application by experts and specialists who can fix the data.

4.1Dataset and Data Preprocessing

We created real-world data set from e-commerce exchange for supervised learning. For the labeled data, we selected items from manufacturers' catalog and assumed most must be correctly spelled. Following table shows labeling of surgical masks manually by experts.

No	Label	Description	Unit of Measurement	Conversion Factor
1	SurgicalMaskSingle	Surgical Mask Disposable	EA	1
2	SurgicalMaskSingle	Surgical Mask Single use	EA	1
3	SurgicalMaskSingle	Surgical Mask use n throw	EA	1
4	SurgicalMask50	Surgical Mask Disposable	BX	50
5	SurgicalMask50	Surgical Mask Single use	BX	50
6	SurgicalMask50	Surgical Mask use n throw	BX	50
7	SurgicalMask50	Surgical Mask Magicare	BX	50
8	SurgicalMask100	Surgical Mask Disposable	BX	100
9	SurgicalMask100	Surgical Mask Single use	BX	100
10	SurgicalMask100	Surgical Mask use n throw	BX	100

Table 4: Labeled Training Data

4.2 Population and Sample

The real-world data for anomaly detection was picked from the invoices submitted by customers. We observed that the normal instances in the dataset are contaminated with anomalies mainly due to erroneous descriptions and typos. Following table shows the anomaly introduced by users for a product called Surgical Mask.

No	Description	Unit of Measurement	Conversion Factor
1	SINGLE USE SURGICAL MASK LEVEL II	EA	1
2	SURGICAL MASK SURGICAL TIE ON	EA1	1
3	SURGICAL MASK ASEPTEX MOLDED	CS	50
4	SURGICAL MASK FLUID RESIS W/ S	BX	50
5	MASK POCKET SURGIC	BX	50
6	MASK,NASAL,HEADSTRP DISP,MEDI/100	BX	100
7	DISPOSABLE SURGICAL PERF. MASK	BX	100
8	MASK SURGICAL POLYPROPYLENE BASIC ELASTIC EARLOOP BLUE	CS	100

Table 5: Surgical Masks with Anomalies

We conducted experiments on real-world health care data sets and tested the performance of our method in terms of detection accuracy, utilization efficiency of labeled data, and robustness to different contamination rates. The experimental results show that the performance of our method is significantly useful to the e-commerce exchange.

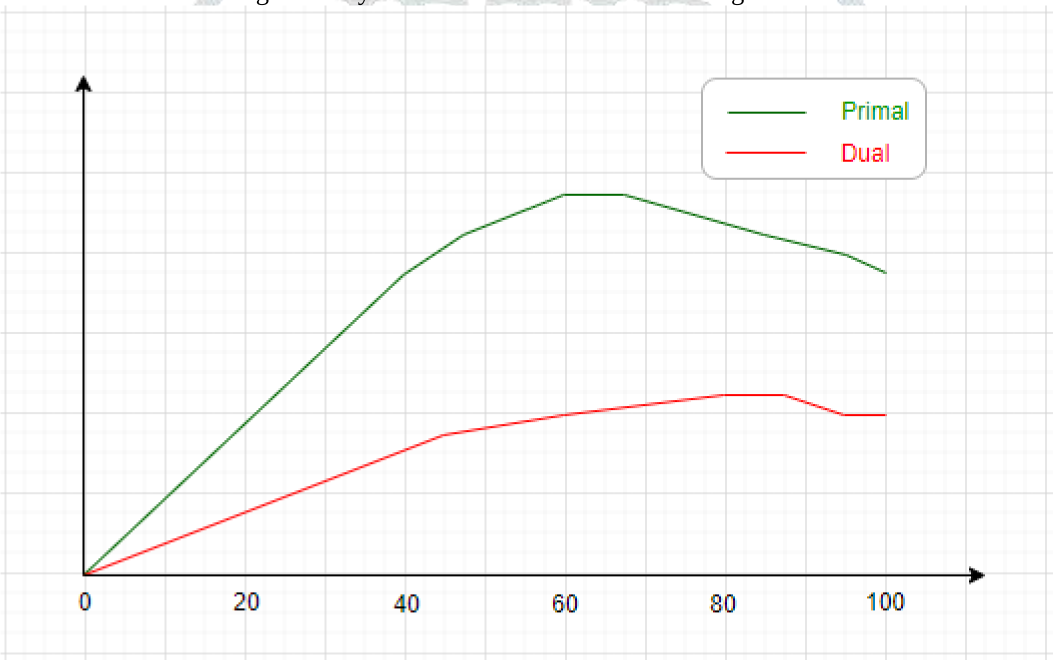


Figure 3: Duality Gap Analysis

The horizontal axis shows the percentage of anomalies in the training set and the vertical axis shows the primal and dual objective values.

4.3 Popular Metrics

In this section, we first describe a set of metrics commonly used for evaluating the performance of our model and then present a quantitative analysis of the performance using popular benchmarks.

Precision, Recall, and F1 Score: These are primary metrics and are more often used for imbalanced test sets. Precision and recall for binary classification are defined in Eq. 1. The F1 score is the harmonic mean of the precision and recall, as in Eq. 1. F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad \textbf{Recall} \quad = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad \textbf{F1-score} = \frac{2 * \mathrm{Prec} * \mathrm{Rec}}{\mathrm{Prec}+\mathrm{Rec}}$$

For multi-class classification problems, we can always compute precision and recall for each class label and analyze the individual performance on class labels or average the values to get the overall precision and recall.

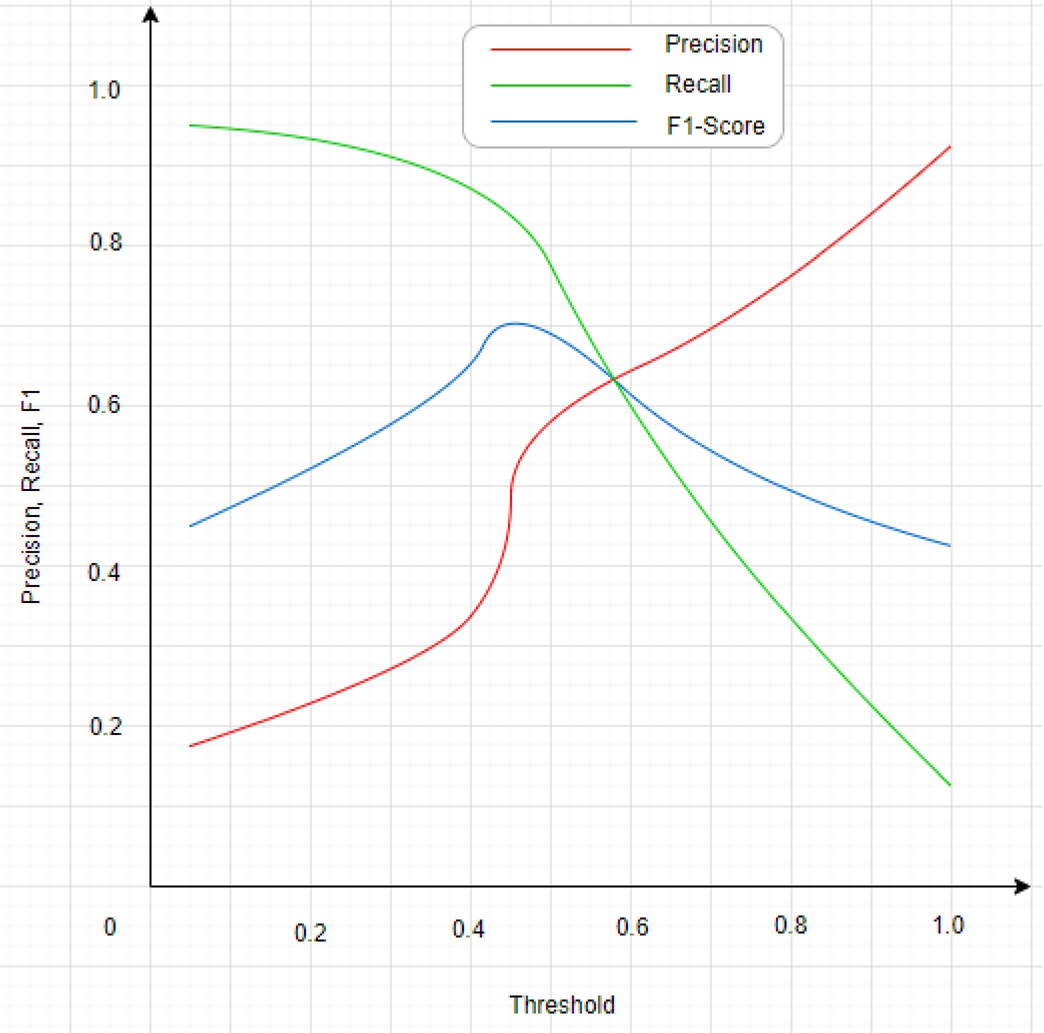


Figure 4: Precision, Recall and F1 curves

From the above results, it can be noticed that better accuracy (as high as F1=7.0) is achieved during our classification with supervised data set of size 10,000. This can be further improved by increasing the labeled data set produced by experts.

4.4 Empirical Results

We aim to use our system to group similar products together using labels. Following table shows the labels assigned to anomalous products explained in Table 5 using our anomaly detection system.

No	Description	Unit of Measurement	Conversion Factor	Label Assigned
1	SINGLE USE SURGICAL MASK LEVEL II	EA	1	SurgicalMaskSingle
2	SURGICAL MASK SURGICAL TIE ON	EA1	1	SurgicalMaskSingle
3	SURGICAL MASK ASEPTEX MOLDED	CS	50	SurgicalMask50
4	SURGICAL MASK FLUID RESIS W/ S	BX	50	SurgicalMask50
5	MASK POCKET SURGIC	BX	50	SurgicalMask50
6	MASK,NASAL,HEADSTRP DISP,MEDI/100	BX	100	SurgicalMask100
7	DISPOSABLE SURGICAL PERF. MASK	BX	100	SurgicalMask100
8	MASK SURGICAL POLYPROPYLENE BASIC ELASTIC EARLOOP BLUE	CS	100	SurgicalMask100

Table 6: Labels assigned using Anomaly Detection System

From the above table, we observe that labels can be assigned to system to remove anomalies. AI-based approach is an alternative to traditional manual processing and cost-effective solution to process millions of records.

V. Conclusion

We believe that incorporating expert knowledge from health care officials will provide valuable insight to this complex process of anomaly detection. There are more than a million products flowing into e-commerce exchange. This data is received on a daily basis, with expectation from decision makers to generate demand insights without huge delay. There are challenges with anomaly detection using human experts like shortage of man power, higher cost, etc. This system is helping us to understand the real-world data and to improve our decision making methods. We use data in real-time, not batches, and learn while simultaneously making predictions. There are no benchmarks to adequately test and score the efficacy of real-time anomaly detectors. While having the ability to predict anomalies is important, it is equally important to be able to guide a human reviewer to the cause of the anomalies. Given that there could be many possible reasons for an anomaly, we need to direct a reviewer to possible suspected issues. In order to concentrate on the most important anomalies to review and in turn further gather labeled data for our models, result of our system is sent to experts for review. Quality audit is conducted at regular intervals to verify the accuracy of our classification. Products identified by experts for training are added to the labeled data set and the system performance in improved. We can further consider more sophisticated learning-based models for explaining anomalies. This will help to prevent the anomalies by educating the users to avoid anomalous description for products.

References

[1] Ramakrishnan, J., Shaabani, E., Li, C., & Sustik, M. A. (2019, July). Anomaly detection for an e-commerce pricing system. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1917-1926).

[2] Weihui Zhu, Xiang Fu and Weihong Han, Online anomaly detection on e-commerce based on variable-length behavior sequence. 11th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2015), 2015, pp. 1-8, doi: 10.1049/cp.2015.

[3] A. Lavin and S. Ahmad, Evaluating Real-Time Anomaly Detection Algorithms -- The Numenta Anomaly Benchmark. 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 2015, pp. 38-44, doi: 10.1109/ICMLA.2015.141.

[4] J. Schneible and A. Lu. Anomaly detection on the edge. MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM), 2017, pp. 678-682.

[5] J. Liang, P. Jacobs, and S. Parthasarathy. Seano: Semi-supervised embedding in attributed networks with outliers. arXiv preprint.

[6] B. Long, Z. Zhang, X. Wu, and P. Yu. Spectral clustering for multi-type relational data. ICML.

[7] Maheshkumar R Sabhnani, Daniel B Neill, and Andrew W Moore. 2005. Detecting anomalous patterns in pharmacy retail data. CMU 2005.

[8] Xianchao Zhang, Jie Mu, Xiaotong Zhang, Han Liu, Linlin Zong, Yuangang Li. Deep anomaly detection with self-supervised learning and adversarial training. Pattern Recognition, Volume 121, January 2022.

[9] Guansong Pang, Anton van den Hengel, Chunhua Shen, Longbing Cao. Toward Deep Supervised Anomaly Detection: Reinforcement Learning from Partially Labeled Anomaly Data. KDD '21: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. August 2021 Pages 1298–1308.

[10] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz et al. Deep Semi-Supervised anomaly detection. Proc. ICLR 2020.

[11] Pang, Guansong, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep Learning for Anomaly Detection: A Review. ACM Computing Surveys (CSUR) 54, no. 2 (2021): 1–38

[12] Xin Kang and Fuji Ren. Understanding blog author’s emotions with hierarchical bayesian models. In 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC), pages 1–6.