**Week1：**

**Mapping:**

1. **Txn -----🡪 word**

**Amt, txnType, receOrpay, chipsOrfed, interbankAmt, JurisdictionAmt, swiftType, etc. time stamp**

**‘Amt\_ txnType\_ receOrpay\_ chipsOrfed\_ interbankAmt\_ JurisdictionAmt\_ swiftType’**

1. **Txn under a single customer -----🡪doc**
2. **Txn group by all customers---🡪all docs--🡪corpus**

**Issue: Why not use structured data but sequence text data?**

**Pre-training downstream**

**Embedding:**

**Pre-train data bought from vendor**

**Frame: W2V, ELMO (bi-direction LSTM), GPT (transformer), BERT (bi-direction, transformer)**

**Above frame is good for detecting sequence pattern which is useful for customer suspicious behavior detection.**

1. **Customer segementaion**

**Txn---🡪(1.0,2.2,3.4)--🡪f(data,dim=3\*1)**

**Txn--🡪balckbox-🡪susp-🡪why? (tagging)**

**23 rule-based scenarios (typology manytoone, onetomany) --🡪file to TM team. False positive rate 98%**

**BERT LSTM-🡪 attention**

**Sigmoid**

**Dense fine-tune**

**Dense fine-tune**

**Dense fine-tune**

**Dense fine-tune**

**Attention fine-tune**

**Time-distribute fixed-par**

**LSTM**

**Embadding**

**Week2:**

**Wire\_txn(swift message, customer CDD information)**

**---🡪 Wire Room (check the format / send RFI**

**and detect the suspicious txn / directly send to TM for investigation) (non-auto: this paper focus on here)**

**---🡪 OFAC screening (auto)**

**---🡪 AML system (auto)**

**---🡪 TM (if txn be alerted by scenarios then will be sent to TM for investigation) (rely on human experience: future work)**

**--🡪 SAR (if customer is determined as true positive, will be filed to FinCEN) (auto)**

**The automation ratio is 40%**

**Future Work: Customer Seg+ Susp Detection**

**Customer Seg:**

**1.NLP approach:**

1. **Hierarchical Dirichlet Process**
2. **NMF (Matrix Decomposition)**
3. **Semi-LDA, Semi- Hierarchical Dirichlet Process**

**2.Normal structure data:**

1. **Clustering, Semi-clustering**

**Susp Detection:**

1. **Multivariate time series Anomaly detection**
2. **NLP Anomaly detection**
3. **NLP text similarity**
4. **GAN can be helpful**

**Week3：**

1. **LDA Cluster--🡪Auto-encoder detection**

**Autoencoder:1. Structure data input train LSTM-auto (benchmark)**

1. **Unstructured data input train auto**

**LSTM-🡪(d1,d2,d3)**

**D1: sample**

**D2: sample-🡪 Sample\_converted**

**rooling\_window=window\_size()**

**D3: feature**

**Sample:**

|  |  |  |
| --- | --- | --- |
| **X0** | **X1** | **X2** |
| **1** | **0.3** | **0** |
| **1** | **0.2** | **1** |
| **1** | **0.3** | **0** |
| **1** | **0.2** | **1** |

**Assue window\_size=2**

**Sample---🡪Sample\_converted**

|  |  |  |
| --- | --- | --- |
| **X0** | **X1** | **X2** |
| **1** | **0.3** | **0** |
| **1** | **0.2** | **1** |

|  |  |  |
| --- | --- | --- |
| **1** | **0.2** | **1** |
| **1** | **0.3** | **0** |

|  |  |  |
| --- | --- | --- |
| **1** | **0.3** | **0** |
| **1** | **0.2** | **1** |

**Task:**

1. **BERT (Attention): Attention**

**LSTM + Attention layer + train**

**BERT(ELMO) pre-train: 2 people**

**Lily, Xiaoxi, gaoyi**

1. **双子模型：Auto-encoder**

**Quro Question pair**

**2 people 双子模型 搭建 （替代Auto-coder）**

**Xingyang, xiaonan**

**3．LDA(变体)： 其他的clustering：2 people**

**Xiaoxi, gaoyi, xiaonan**

**Lit review: Kecheng Xingyang**