Identifying Vendors using Topological Data Analysis Methods

MSc Data Science, Newcastle University

Deepika Chandramouli, Mahbub Gani, Tong Xin

INTRODUCTION

- TDA excels in offering deep insights into the topological structures of large and intricate datasets, with successful applications in fields like cancer research, material science, and 3D shape analysis [1].
- This research leverages TDA techniques on bounding box information from OCR-scanned receipts to enhance vendor identification process.
- A **vendor** is the entity from which goods or services are purchased (e.g., Walmart, Aldi), and accurately identifying these vendors from receipts is crucial for financial tracking and analysis.
- By combining TDA with machine learning, particularly in collaboration with financial software like Sage, the study aims to overcome existing limitations such as inefficient data processing and error-prone vendor identification

MOTIVATION

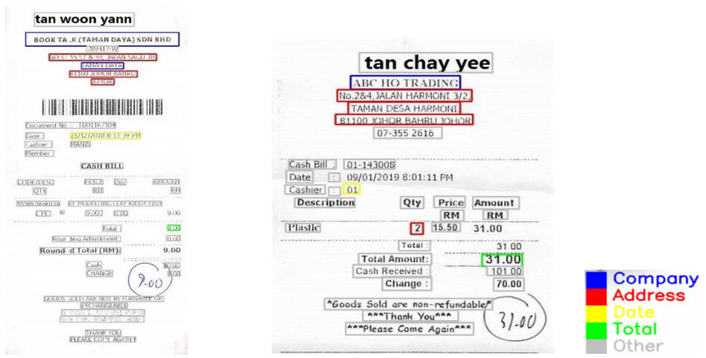
When and where TDA is effective?

- **Data Heterogeneity**: Financial documents vary widely in format and structure. Traditional data analysis methods often struggle to handle this heterogeneity, leading to inefficiencies and errors.
- **Improved risk management**: TDA enhances risk management by uncovering anomalies and unusual patterns in vendor data, enabling proactive identification and mitigation of potential issues before they escalate.
- **Optimisation of Vendor Selection:** TDA optimises vendor selection by analysing attributes and performance metrics to identify the most suitable vendors and better align them with organisational goals.

OPTICAL CHARACTER RECOGNITION

Leveraging bounding box coordinates from OCR can offer:

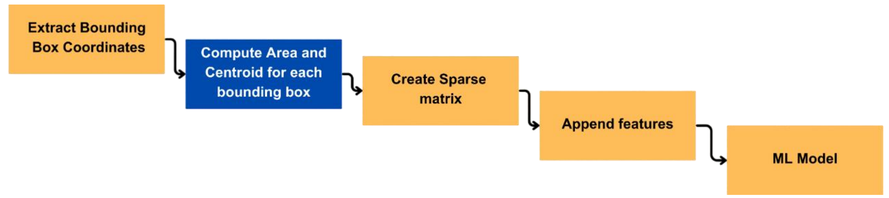
- Use bounding box coordinates to accurately extract vendor details and segment documents into sections (e.g., headers, body text) for improved data processing and easier navigation.
- Analyse metadata to compare vendor information across documents, aiding in better decision-making and integration with vendor management systems.



METHODS

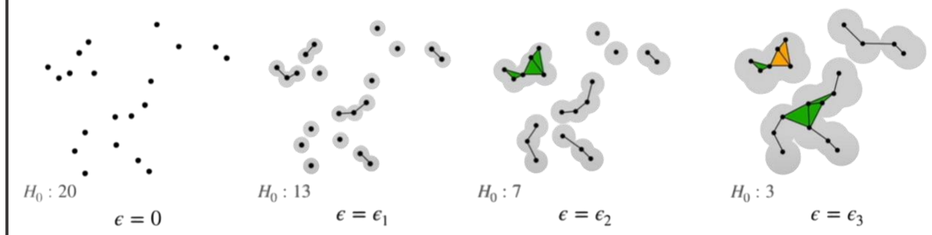
1. Baseline model:

- **Data Processing:** Extract bounding box coordinates from the SROIE dataset [2], compute areas and centroids, and convert this information into a feature matrix using a sparse representation.
- **Model Training:** Train any ML classification model such as Random Forest classifier using the processed feature vectors from the training dataset. (Other ML models are also performed)
- **Evaluation:** Predict vendor identities on the test dataset and calculate the model's accuracy to assess performance.

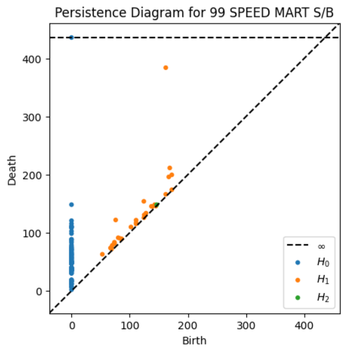


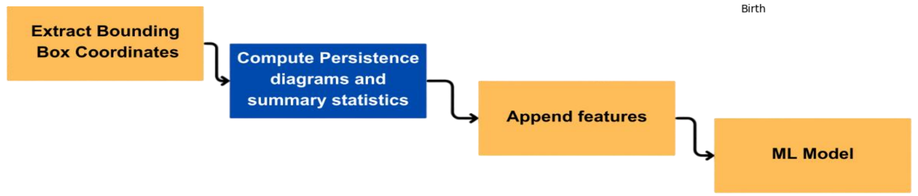
2. Topological data analysis method:

- **Persistent Homology** tracks the evolution of topological features (connected components, loops, voids) across multiple scales, identifying features that persist and are significant versus those that are likely noise [3].



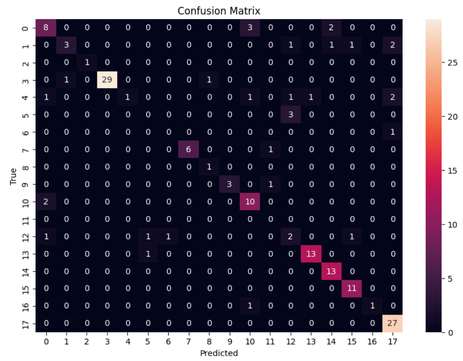
- **Persistence Diagrams** are used to capture the birth and death of topological features (e.g., clusters, loops) in data as scale varies.
- Use bounding boxes to understand data spread and noise by analysing the persistence of these features.
- Features close to the diagonal are considered noise, as they appear and disappear quickly, while features farther from the diagonal exhibit long persistence and represent meaningful data structure [3].
- Summary statistics from the persistence diagrams are used to create a feature vector for ML model application.





RESULTS

- Baseline approach serves as a solid benchmark; However, it primarily focused on numerical features and lacked the ability to uncover complex data patterns.
- The TDA model achieved 80% accuracy, slightly lower than the baseline machine learning model but demonstrated TDA's strength in identifying anomalies and managing heterogeneous data.
- The TDA approach uncovers complex topological structures, such as clusters and loops, reveals the persistence of these features across scales, and identifies anomalies, providing a deeper understanding of data relationships and patterns that complement traditional methods.



Accuracy of Random Forest Classifier	
Baseline Model	TDA model
0.95	0.80

```
Actual: GERBANG ALAF RESTAURANTS SDN BHD, Predicted: GERBANG ALAF RESTAURANTS SDN BHD
Actual: TEO HENG STATIONERY & BOOKS, Predicted: AEON CO. (M) BHD
Actual: TEO HENG STATIONERY & BOOKS, Predicted: SYARIKAT PERNIAGAAN GIN KEE
Actual: UNIHAKKA INTERNATIONAL SDN BHD, Predicted: UNIHAKKA INTERNATIONAL SDN BHD
Actual: UNIHAKKA INTERNATIONAL SDN BHD, Predicted: UNIHAKKA INTERNATIONAL SDN BHD
Actual: UNIHAKKA INTERNATIONAL SDN BHD, Predicted: UNIHAKKA INTERNATIONAL SDN BHD
```

CONCLUSION

- Expand the combined TDA and machine learning framework to various document types and industries, leveraging TDA insights to enhance model performance and address risks, fraud, and errors in vendor management.
- Implement the integrated approach with existing data processing and financial software like Sage to create a more efficient and comprehensive data analysis workflow, enhancing overall decision-making capabilities.
- **Future Work**: Further optimisation, addressing class imbalance, and integrating TDA with machine learning, along with data augmentation techniques, could significantly enhance the accuracy and efficiency of vendor identification systems.

REFERENCES

[1] Edelsbrunner, Letscher & Zomorodian Topological Persistence and Simplification. Discrete Comput Geom 28, 511–533 (2002).

[2] SROIE v2 dataset from Kaggle https://www.kaggle.com/datasets/urbikn/sroiedatasetv2 - Introduced by Huang et al. in ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction

[3] https://medium.datadriveninvestor.com/persistent-homology-f22789d753c4

CONTACT

d.chandramouli2@newcastle.ac.uk