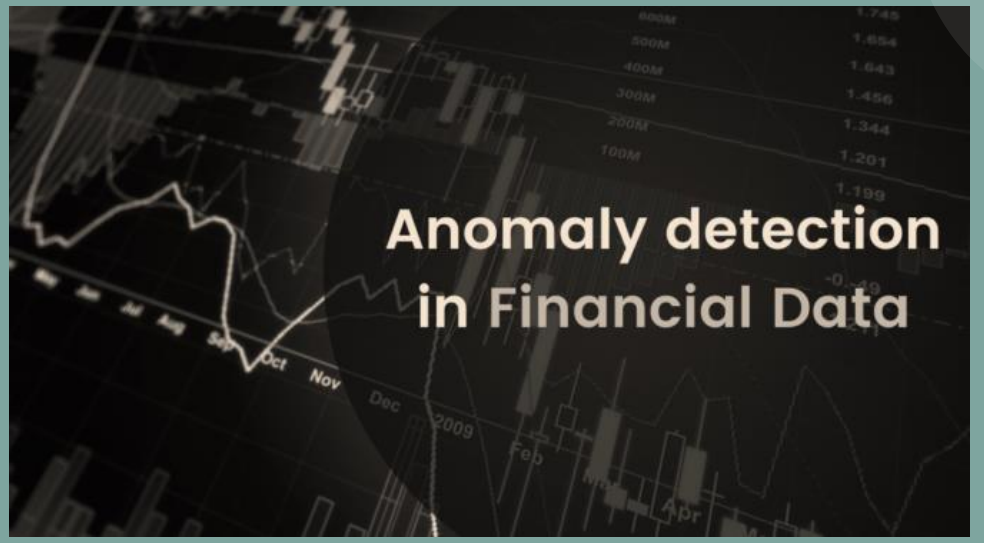
Textual Anomaly Detection in Financial Reports

DSCI 6004: Natural Language Processing

1. Roshini Bandi
2. Prasad Thamada

Project Objectives

- The objective of this project is to develop an automated system for detecting textual anomalies in financial reports through advanced NLP techniques.
- The system aims to enhance the accuracy and efficiency of anomaly detection in financial documents, thereby assisting financial analysts and auditors in identifying potential fraud or errors



Statement of Value

THIS PROJECT HOLDS SIGNIFICANT VALUE IN THE FINANCIAL DOMAIN BY ADDRESSING THE CRITICAL NEED FOR ACCURATE ANOMALY DETECTION IN FINANCIAL REPORTS. LEVERAGING ADVANCED NATURAL LANGUAGE PROCESSING (NLP) METHODS, IT AIMS TO ENHANCE DETECTION PROCESSES, ENSURING DATA INTEGRITY, TRANSPARENCY, AND TRUST. BY AUTOMATING ANOMALY DETECTION, THE PROJECT NOT ONLY IMPROVES ACCURACY AND SPEEDS UP ANALYSIS BUT ALSO CONTRIBUTES TO COST SAVINGS AND RISK MITIGATION BY IDENTIFYING IRREGULARITIES EARLY.

THE OUTCOMES EXTEND BEYOND IMMEDIATE APPLICATIONS, PAVING THE WAY FOR ADVANCEMENTS IN FINANCIAL TEXT ANALYSIS, REGULATORY COMPLIANCE, AND PREDICTIVE ANALYTICS, THUS INFLUENCING INDUSTRY PRACTICES AND GOVERNANCE FRAMEWORKS. THIS PROJECT ALIGNS WITH INDUSTRY IMPERATIVES FOR RESPONSIBLE DATA STEWARDSHIP, MAKING IT A COMPELLING AND IMPACTFUL ENDEAVOR IN FINANCIAL TECHNOLOGY AND GOVERNANCE.

State of the Art

Barrett et al. [2] aim to minimize human intervention in log file processing by proposing a novel approach that treats logs as regular text, departing from previous methods focused on exploiting the limited structure imposed by log formatting. Our approach leverages modern natural language processing techniques, starting with the application of a word embedding method based on Google's word2vec algorithm. This technique maps words from log files into a high-dimensional metric space, which we then utilize as a feature space with standard classifiers. The resulting pipeline is highly versatile, computationally efficient, and requires minimal manual intervention. To validate our approach, we conducted experiments to identify stress patterns on an experimental platform. The results demonstrate strong predictive performance, achieving approximately 90% accuracy, using three readily available classifiers.

State of the Art

The study by Bertero et al. [2] addresses the complex challenge of identifying textual outliers within Risk Factors sections extracted from a vast corpus of U.S. corporate annual reports. By employing recent outlier detection methods tailored for high-dimensional textual data, including LOF+PCA, TONMF, and KNN algorithms, the research aims to uncover anomalies in language usage, such as topic deviations and stylistic variations. The experimental evaluation involved analyzing 150,000 annual reports without preprocessing text, using a basic unigram-based bag-of-words feature set. Results indicate that the TONMF model outperformed others with an F1 score of .62 and an MCC score of .47, showcasing higher precision in outlier detection. However, despite significant p-values in Fisher's Exact Test for TONMF and LOF+PCA, the overall Krippendorff's alpha score suggests room for improvement in making more human-like selections, possibly through additional features capturing semantic nuances and financial contexts.

Approach

1. **Preprocessing:** Clean and preprocess the textual data to remove noise, standardize the format, and handle missing values.
2. **Feature Extraction:** Utilize NLP techniques such as word embeddings, syntax analysis, and semantic parsing to extract meaningful features from the text, including key terms, entities, and relationships.
3. **Anomaly Detection Models:** Develop and train machine learning models, including unsupervised anomaly detection algorithms such as Isolation Forestand One-Class SVM) and supervised classifiers such as Random Forest, LSTM), using the extracted features.
4. **Tools and Techniques:** Employ Python programming language, NLP libraries machine learning frameworks such as TensorFlow, PyTorch, and relevant APIs for data processing, model development, and evaluation.
5. **Datasets:** Gather a diverse dataset of financial reports from Kaggle (https://www.kaggle.com/datasets/shashankkumarranjan/financial-anomaly-data) and (https://www.kaggle.com/datasets/mustafakes) for training and testing the models.
6. **Algorithms:** Experiment with various algorithms and techniques, such as ensemble methods, deep learning architectures and anomaly scoring methods, to enhance anomaly detection accuracy and robustness.

Deliverables

Preprocessing and feature extraction:

The project includes robust preprocessing and feature extraction pipelines to ensure data quality and relevance for anomaly detection. This involves tasks such as data cleaning, normalization, tokenization, and feature engineering to transform raw financial text into structured input suitable for machine learning models.

Anomaly Detection Model Training:

The focus is on developing anomaly detection models that achieve high accuracy and interpretability. This involves selecting appropriate algorithms such as supervised or unsupervised learning techniques, considering the nature of financial data, and optimizing model parameters to detect anomalies effectively while minimizing false positives.

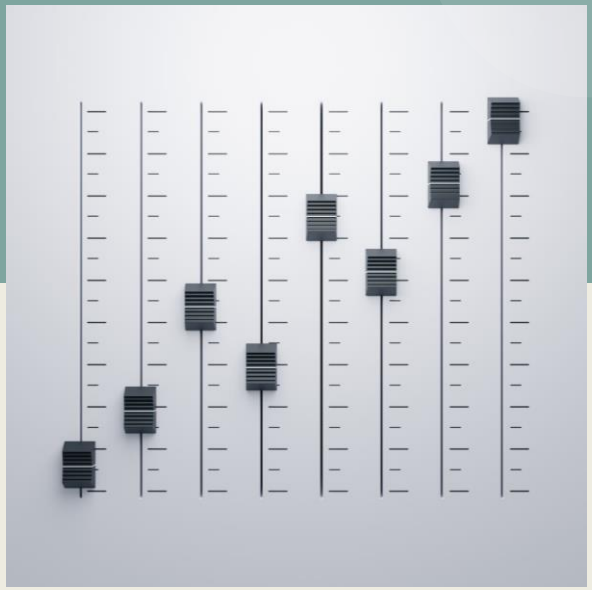
Performance Evaluation and Analysis:

A comprehensive performance evaluation report will be generated, providing detailed analysis using key metrics such as precision, recall, F1-score, and others relevant to anomaly detection tasks. The report will also include a comparison against baseline methods or existing systems to demonstrate the effectiveness of the developed models.

Documentation and User Guide:

This documentation will cover deployment recommendations, configuration settings, and best practices for utilizing the anomaly detection models in real-world financial environments.

Evaluation Methodology



1. **Metrics:** We will use standard evaluation metrics such as precision, recall, F1-score, and accuracy to quantify the performance of our anomaly detection models. These metrics will provide insights into the models' ability to correctly identify anomalies while minimizing false positives and negatives.
2. **Validation Techniques**: Cross-validation methods will be utilized to validate the generalization capabilities of our models. By splitting the dataset into training and testing sets multiple times and evaluating performance across different folds, we can ensure that the models perform consistently across varying data samples and avoid overfitting.
3. **Comparison with Baselines:** Our system will be benchmarked against existing anomaly detection methods and traditional rule-based approaches commonly used in financial analysis. This comparative analysis will highlight the superiority of our NLP-based approach in detecting textual anomalies within financial reports, showcasing its enhanced accuracy, interpretability, and efficiency.
4. **Qualitative Analysis:** we will conduct qualitative analysis by reviewing specific examples of detected anomalies. This qualitative assessment will provide deeper insights into the types of anomalies identified by our system and its ability to capture nuanced irregularities in financial text data.

Cited Work

1. Barrett, L., Fletcher, S., Ortan, A., & Kingan, R. (2019, October). Textual Outlier Detection and Anomalies in Financial Reporting. In Proceedings of the 2nd KDD Workshop on Anomaly Detection in Finance (pp. 1-10). Anchorage, Alaska, USA: Bloomberg BNA.
2. Bertero, C., Roy, M., Sauvanaud, C., & Tredan, G. (2017, October). Experience Report: Log Mining Using Natural Language Processing and Application to Anomaly Detection. In Proceedings of the 2017 IEEE 28th International Symposium on Software Reliability Engineering (ISSRE) (pp. 1-8). doi:10.1109/ISSRE.2017.43

Thankyou