Transport and Telecommunication Institute

Faculty of Engineering Science

Artificial Intelligence Group Project

Project proposal

"ADVERSE MEDIA MONITORING AND CLIENT RISK ASSESSMENT SYSTEM"

Students:
Natalja Krjuckova ST50446
Sergejs Kopils ST83519
Agita Ferstere ST84950
Diāna Koržaviha ST58392

Study Group: 4303MDA

RIGA 2024

Contents

1. Main problem

With the growing importance of monitoring client activities, to improve the identification of customer problems, automated risk assessment has become essential. For the purpose of helping financial institutions better detect possible problems with their clients, this project intends to create a system for tracking adverse media coverage and evaluating client risk. By looking through publicly available negative information, such as news articles, court documents, regulatory filings, and social media, the technology will automatically assess clients and group them into risk groups. This method aims to give a deeper look at consumer data while reducing manual work. This strategy will assist protect institutions' reputations and reduce their susceptibility to financial crimes by guaranteeing adherence to Know Your Customer (KYC) and Anti-Money Laundering (AML) regulations.

2. Initial research on the problem domain

Previous investigations on adverse media monitoring and customer risk evaluation underscore the significance of Natural Language Processing (NLP) and Artificial Intelligence (AI) in enhancing efficiency and precision (Banerjee and Roy, 2022). Existing products mainly use the existing library of media resources and infrequently incorporate hybrid machine-learning approaches that combine supervised and unsupervised models for adaptive risk profiling. This results in a lack of sensitivity to subject-specific context and limited information on emerging risks (Emmanuel Agwu, 2023).

3. Project objectives and scope

The objectives of this project are:

- to implement data aggregation techniques to assess client risks by analysing adverse media content;
- to implement machine learning models for the identification of adverse media related to financial crimes for client risk classification;
- to assess the ability of chosen algorithms, to classify clients into different risk categories.

Optional:

- the supervised model could be a second part of the project.

Out of Scope:

- transaction monitoring of client;
- closed-source private databases or proprietary client information.

4. Proposed work activities

It is expected that tools, frameworks and libraries will be chosen depending on the test results of each method.

1. Automatic collection of customer references in media (Web Scraping) – it is planned to use such tools, frameworks and libraries: Selenium, BeautifulSoup, Scrapy, Airflow and API for sources, that provide API for accessing publications. The following list of subtasks should be completed in this part:
   1.1. defining resources for data collecting;
   1.2. keywords list creation for content filtering based on company names;
   1.3. automatic scripts are developing and configuring, which includes requesting, parsing HTML and extracting data, keyword filtering, saving data and scheduling regular data collections.
2. Classification of collected data for risk assessment:
   2.1. collected data pre-processing – converting text into a clean sequence of keywords, ready for vectorisation;
   2.2. converting text to vector data – converting data into a numerical representation using tf-idf, word2vec or glove so that clustering algorithms can process them;
   2.3. clustering using k-means or DBSCAN algorithms;
   2.4. analysis and cluster interpretation - to determine which groups of articles may be associated with risk. cluster interpretation can be conducted by analysis of keywords and topics in each cluster, manual verification of articles and assignment of risk levels;
   2.5. cluster evaluation and improvement – that includes checking the clustering results and anomalies, the number of clusters increasing or decreasing will be performed if necessary;
   2.6. final evaluation and report - after clustering and interpreting the clusters, a descriptive report should be prepared describing each cluster and its risk level (high, medium, low);
   2.7. assigning a risk level to each client—identifying the appropriate risk level based on their association with clusters.
3. Supervised model training based on the results (Optional).
4. Model evaluation (Precision, Recall, F1-score) and testing (Optional).
5. Result analysis and generating a risk report
   5.1. Based on modelling, a report for each client will be created that includes the number and content of articles associated with the client and its risk level (high, medium, low).

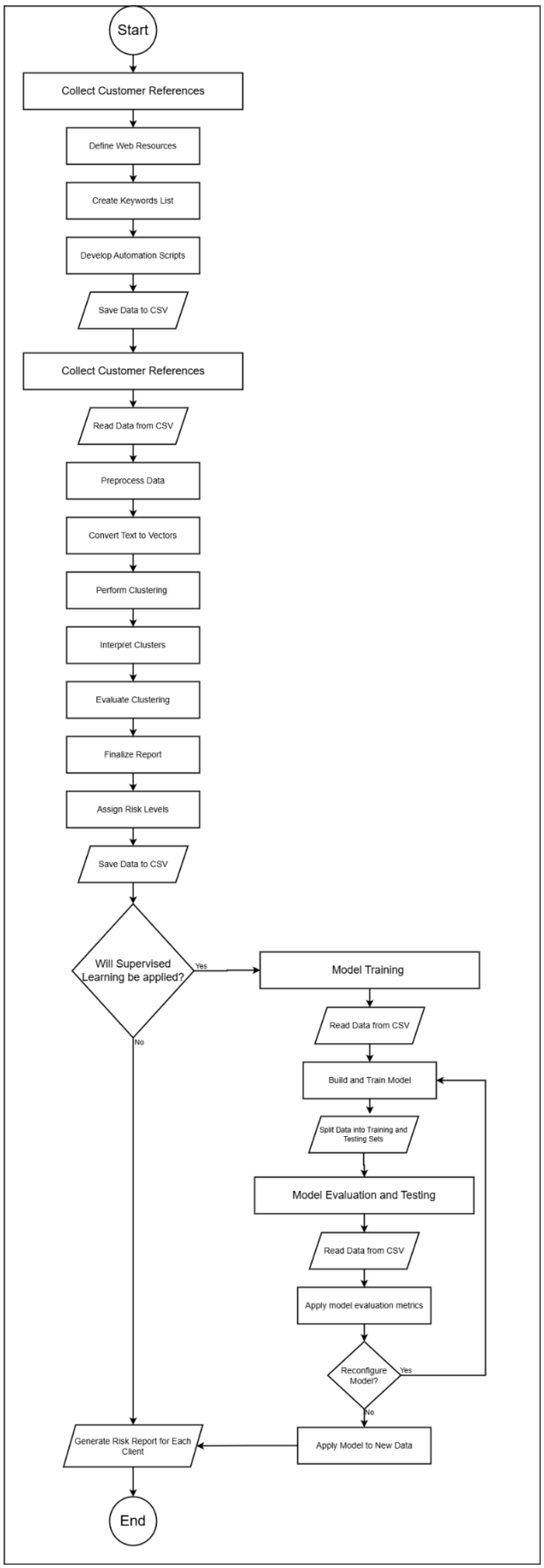

Figure 1 Workflow diagram

5. Project management methodology

This project will use an Agile project management approach with a Kanban project management framework. GitHub's planning and tracking tools would be used to manage projects with the support of regular Teams calls as an additional communication method. Following the project meeting, a post-meeting memo will be distributed, highlighting the meeting's key points, decisions made, and any adjustments to sprint objectives. Link to GitHub - https://github.com/NataKrj/AI-project-2024.

6. Project team

Project Manager: Sergejs Kopils

Knowledge about Agile approaches to guarantee timely, effective task completion. Excellent presenting abilities based on the architecture degree. For educational reasons, field-related data science experience should be less than one year.

Lead Developer: Diāna Koržaviha

IT work experience > 5 years, experience in web scraping technologies, Python for data analysis, integrating APIs and applying machine learning methods to data scheduling and text analysis.

Developer 1: Natalja Krjuckova

Junior developer role, experience in data science with application to everyday work activities. Expert-level knowledge of the business-related issue for a particular project directly contributing to overall project objective attainment.

Developer 2: Agita Ferstere

Position as a junior developer, with less than a year of field experience in data science, primarily for educational purposes. Basic familiarity with clustering techniques to aid in execution.

	Critical Activity	Project Manager	Lead Developer	Developer 1	Developer 2
Administration	Schedule and conduct group meetings	R	C	C	C
	Project implementation artefact collection and recording	R	C	C	C
	Establish priorities for ongoing support, maintenance, critical issues and enhancement of the application	A	R	A	A
	Manage scope, resources and issues related to project management	R	R	A	A
	Project visual design and presentation preparation	R	A	A	A
	Project proposal	A	R	A	A
Implementation process	Defining Resources for Data Collection	A	C	R	C
	Keywords List Creation	C	R	A	C
	Automatic Script Development and Scheduling	I	R	C	C
	Classification of Collected Data for Risk Assessment	I	C	R	C
	Data Pre-processing and Vectorization	I	R	A	C
	Clustering and Interpretation	C	R	A	A
	Evaluation	A	R	A	A

KEY		
	R = Responsible	Responsible for performing the task (ie. the actual person doing the work to complete the task).
	A = Accountable	Ultimately accountable for the task being done satisfactorily. The accountable person must sign-off the work that the Responsible person produces.
	C = Consulted	Team members whose input is used to complete the task. Communication with these members will be 2-way in nature.
	I = Informed	Team members who are informed as to the status of the task. Communication with these members will be 1-way in nature.

Figure 2 RACI Matrix

REFERENCES

1. Banerjee, P. and Roy, R. (2022) Integrating Natural Language Processing (NLP) in AML Compliance and Monitoring. 11.

2. Emmanuel Agwu (2023) *A Practical Guide to Adverse Media Screening: Best Practices and Tools Youverify website*. 3 August 2023 [online]. Available from: https://youverify.co/blog/practical-guide-to-adverse-media-screening [Accessed 23 October 2024].

3. Adverse media or negative news screening (2024) sanction scanner. Available at: https://www.sanctionscanner.com/knowledge-base/adverse-media-144 [Accessed: 27 October 2024].

4. Ashtikar, A. (2024) Applying vector databases in finance for risk and fraud analysis, Zilliz blog. Available at: https://zilliz.com/learn/applying-vector-databases-in-finance-for-risk-and-fraud-analysis [Accessed: 27 October 2024].

5. Pavanbelagatti (2024) Pavanbelagatti/vector-embeddings, GitHub. Available at: https://github.com/pavanbelagatti/vector-embeddings [Accessed: 27 October 2024].

6. ProjectPro, (2024) 8 feature engineering techniques for Machine Learning, ProjectPro. Available at: https://www.projectpro.io/article/8-feature-engineering-techniques-for-machine-learning/423 [Accessed: 27 October 2024].

7. Tirumalachandraveni (2024) Fake news detection model using tensorflow in python, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/fake-news-detection-model-using-tensorflow-in-python/ [Accessed: 27 October 2024].

8. Vaishnavi, A. (2024) NIFTY50 companies ESG score data, Kaggle. Available at: https://www.kaggle.com/datasets/akulvaishnavi/nifty50-companies-esg-score-data [Accessed: 27 October 2024].

9. Wu, J. (2022) Web scraping with Python: Automate negative news screening (NNS) at the internet search engine, Medium. Available at: https://medium.com/@jasonclwu/web-scraping-with-python-automate-negative-news-screening-nns-at-internet-search-engine-c99697080b14 [Accessed: 27 October 2024].