AML/KYC team: Automating Risknews for AML/KYC using LLM

Executive Summary/Project Description

**Problem Statement**

The AML/KYC team currently uses the Transparint tool for monitoring and ad-hoc research but wants to replace it with a more efficient tool, ComplianceNews. The goal is to monitor a list of third parties and search for **negative articles from an AML/KYC perspective**.We intend to leverage from another similar project, RiskNews, which measure the context, mood and tone of news articles and produce scorings on articles and GEMS’s couterparties to assess credit risk.

**Solution Overview**

We propose **automating the monitoring and article search** process using an advanced language model (**LLM**). The LLM will measure the context, mood and tone of news articles and produce **scorings** on articles and **AML/KYC** related people (counterparties, physical people, shareholders, stakeholders, politically exposed people). These scorings will be used to assess AML/KYC risk and generate accurate reports. By leveraging LLM technology, we aim to improve efficiency, consistency, and risk assessment. In addition, as an optional feature to improve scoring, we can use **World-Check** feed, a comprehensive database and service that helps financial institutions and companies comply with KYC, AML, CFT, anti-bribery, and corruption regulations by providing accurate and reliable information from reputable sources. World-Check will help us identify stakeholders in articles with potential AML/KYC risk and flag them for further review.

**Key Features**

* **ComplianceNews UI**:
  + Developed by the Risk Digital team.
  + Allows users to monitor and search for negative AML/KYC articles.
* **LLM Implementation**:
  + The LLM will replace manual report generation.
  + Analyses articles and identifies relevant topics for AML/KYC.
  + Measures the context, mood, and tone of news articles.
  + Produces scorings on articles, counterparties, and a list of physical people, shareholders, stakeholders, and politically exposed people.

**Benefits**

* **Efficiency**: The LLM reduces manual effort, allowing analysts to focus on higher-value tasks.
* **Consistency**: Standardized reports ensure uniformity in assessments.
* **Risk Reduction**: Rapid identification of relevant articles minimizes compliance risks.

**Next Steps**

We seek budget approval to implement the LLM solution, which promises significant gains in productivity and risk management.

Use Case Description

**Goal**

The primary objective of this project is to leverage generative AI technologies to automate the monitoring and search for negative articles for AML/KYC.

**Technical Approach**

1. **Relevant articles/AML/KYC counterparties matching**
   * **Article Retrieval and Storage**
     + Implemented a pipeline to retrieve articles from Lexis Nexis
     + Implement a pipeline using an api to feed from World Check database (api: <https://developers.lseg.com/en/api-catalog/customer-and-third-party-screening/world-check-one-api>)
   * **Extracting Accurate Information**:
     + Filter articles to produce the AML/KYC scope.
   * **AML/KYC counterparties**:
     + **Companies matching**:
       - Extract a list of counterparties from AML/KYC perspective from GEMFORCE and other internal sources
       - Clean and preprocess counterparties to improve matching.
       - Enrich counterparties scoring with World Check database
       - matching process between mentioned companies in the news and AML/KYC counterparties
2. **Pure LLM**
   * **Converting Information into Sentences**:
     + Build prompts to guide the LLM towards generating accurate scoring.
     + Use few-shot learning to illustrate the desired scoring.
3. **Model Validation**
   * **Qualitative Validation (Feedback)**:
     + Present LLM-generated scoring to stakeholders.
     + Gather feedback and adjust the model if necessary.
     + Regular tests will be conducted by AML/KYC analysts to improve the model's performance.
4. **Integration**
   * **Integrate the LLM into Client Interfaces**:
     + **Automate** **scoring** generation in the ComplianceNews interface.
     + include tabs for article/ AML/KYC counterparty view, filters by score, count number of articles/ AML/KYC counterparties on tables, and a **search** function for article/counterparty.
     + Created a **heatmap** for each sector/sub-sector scores. The score heatmap is based on underlying AML/KYC counterparties.
     + Provide real-time insights to AML/KYC analysts.

Expected Results

The LLM-based system will:

* Analyze articles and identify relevant topics for AML/KYC.
* Identify counterparties in the articles and score their AML/KYC risk.
* Generate precise and contextual scoring.
* Provide actionable insights to analysts for adjusting their assessments.

**Quantifiable Benefits**

* **Time Savings**: Reduced manual effort for report generation.
* **Accuracy**: Consistent and reliable report generation.
* **Risk Reduction**: Rapid identification of relevant articles.

**Business Impact**

* **Faster Response to AML/KYC risk**.
* **Improved Communication between Analysts and Stakeholders**.

Key Success Factors

* **Quality of Generated Reports**:
  + Ensure that the LLM produces relevant scoring or articles and properly identify AML/KYC related stakeholders.
* **Root Cause Identification**:
  + Correctly identify the underlying causes of negative articles.

Data Used

* **Raw Data**:
  + **LexisNexis**: For retrieving news **articles** related to the AML/KYC counterparties.
  + **World Check**: As a potential additional source for **counterparties matching** (<https://developers.lseg.com/en/api-catalog/customer-and-third-party-screening/world-check-one-api>)
  + **GEMFORCE**: For monitoring a list of **third parties**.
  + **Internal Resources**: for additional list of **third parties**.
* **Storage**:
  + **Articles stored in AWS S3** and partitioned by year, month, and day.
  + **Counterparties enriched with World Check stored in AWS S3** and partitioned by year, month, and day.

Technologies Available

* **OpenAI API**:
  + Access to the API in the Azure resource group.
* **GPT 3.5 Turbo/GPT 4**:
  + Allows asking questions and analysing data in natural language.
* **Text Embedding Techniques:**
  + It might be required to convert articles upstream into numerical vectors for faster semantic understanding of text data.
* **Model Training:** 
  + tailor the model to better understand and respond to the prompts

Team

* **Digital Risk Team**: project development / project management
  + Steven Roland (Business Analyst/Project Manager)
  + Hakim Boulahya (Python Developer lead/DevOps)
  + Sebastien Laloo (Python Developer/Data Scientist)
  + Mathias Merdjan (Python Developer/DevOps)
* **AML/KYC Risk**: sponsor
  + Martial de Saint Loubert (Credit Risk Team Leader)
  + Menna Glabal (Credit Risk Controller)

Budget

### **Expertise and Synergies**

We possess expertise in the technologies required for this project. Our team has successfully utilized similar technologies in previous projects, including:

1. **Newspy:**
   * Utilized OpenAI for context-aware analysis of news articles for scoring purpose
   * Developed an automated system to assess credit risk signals (PD) for GEMS clients.
   * Budget: Initially allocated €50,000, which was ultimately exceeded due to project complexity.
2. **Market Data Monitoring:**
   * Leveraged machine learning algorithms to detect and flag market data anomalies.
   * Created a user interface for risk officers and traders to validate and correct errors.
   * Budget: Consumed €150,000.

**Estimated Development Costs:**

Based on our experience with previous similar projects and the project steps outlined in our use case, we provide the following cost estimates:

* + Counterparties Matching: 4 weeks, €35,000
  + Pure LLM: 4 weeks, €35,000
  + Model Validation: 5 weeks, €45,000
  + Integration: 3 weeks, €30,000
  + **Total**: €145,000

We estimate a total development cost of **€145,000**.

By leveraging our experience and optimizing synergies, we aim to deliver a successful project within the allocated budget

**Estimated Prompt Token Costs:**

We have a collection of around **20,000 articles of rank 1**, which means that they are articles of best quality in terms of relevance, accuracy, and clarity. The quality of articles is essential to ensure that our LLM model can provide correct and reliable answers to scoring legal entities.

**Assuming** the project requires to analyse around **20000 articles** per day:

1. **Project Description:**

* **Articles per day:** 20,000
* **Max tokens per request:** 4,096 with 1 request per article
* **Total tokens per day:** 20,000 \* 4,096 = 81,920,000 tokens
* **Price per 1,000 tokens:** €0.001 (for GPT-3.5-Turbo)

1. **Daily Cost Calculation:**

Daily Cost=(81,920,000 tokens / 1,000) ​×€0.001=€81.92 per day

1. **Monthly Cost Calculation:** Assuming 21 business days in a month:

Monthly Cost=€81.92 per day×21 days=€1720.32 per month

**Summary**

* **Daily Cost:** €81.92
* **Monthly Cost:** €1720.32

The project variable costs is directly related to the volume of articles processed daily. By optimizing token usage, filtering higher quality articles and implementing cost management strategies, you can potentially reduce these costs.

**Estimated World Check api Costs:**

For each legal entity, the number of people to screen with World-Check to identify potential AML/KYC risks or compliance issues can vary significantly depending on the size and nature of the company, as well as the industry it operates in. However, at first, let’s assume we will only focus on **main key stakeholders** of each company with an average of **10** **people per company**:

**1. Project Description:**

* **Clients**: 15,356
* **Key stakeholders per client**: Assuming an average of 10 key stakeholders per client
* **Total screenings per day**: 15,356 clients \* 10 stakeholders = 153,560 screenings
* **Price per screening (credit)**: $0.10

**2. Daily Cost Calculation:**

* **Daily Cost**: 153,560 screenings \* $0.10 per screening = $15,356 per day

**3. Monthly Cost Calculation:**

Assuming 21 business days in a month:

* **Monthly Cost**: $15,356 per day \* 21 days = $322,476 per month

**Summary:**

* **Daily Cost**: $15,356
* **Monthly Cost**: $322,476

The screening could be extended to all companies’ customers or, for high-risk industries, to all employees and business partners but let’s assume key stakeholders will already provide a good rating of the companies’ AML/KYC risk profile.

Conclusion

Our project aims to automate the monitoring and search for negative articles for the AML/KYC team. By using advanced language models (LLM) techniques, we will improve risk assessment and decision-making. Here are the key points:

* **Objective**: Automate the generation of reports on negative articles on a AML/KYC perspective.
* **Approach**: Use an LLM to analyse data and generate contextual reports and scoring.
* **Budget**: Estimated **development** cost of **€145,000**, **monthly openai** API cost of **€1720.32** and monthly api cost of **€322.47** forWorld Check if required

In summary, our project promises efficiency gains, risk reduction, and improved communication for the AML/KYC team. Let’s proceed with confidence, leveraging our expertise and maximizing synergies with Risknews developed for credit risk analysis purpose.