KNN Classification Model for dual Search Intent categories predictions with explainable AI in Healthcare sector of French market

Phool Kumar, Assistant Professor - IAE - University School of Management, University of Lille - France, phool.kumar@univ-lille.fr
Dominique Crie, Full Professor - IAE - University School of Management, University of Lille - France, dominique.crie@univ-lille.fr
Annabel-martin Salerno, Full Professor - IAE - University School of Management, University of Lille - France, annabel.salerno@univ-lille.fr
Hamdi Dkhil, Associate Professor - IAE - University School of Management, University of Lille - France, hamdi.dkhil@univ-lille.fr
Mihai Calciu, Professor Emeritus - IAE - University School of Management, University of Lille - France, mihai.calciu@univ-lille.fr

To be cited as: Kumar, P., Crie, D., Salerno, A. Dkhil, H., & Calciu M. (2025). KNN classification model for dual search intent categories predictions with explainable AI in healthcare sector of French market. In *Proceedings of the IMTC Conference Berlin 16-17 Jan 2026* (pp. **[Page Range]**). Paris-Venice Marketing Trends Association. **[DOI]**.

Abstract - The growing reliance on digital platforms for healthcare information has underscored the importance of accurately identifying user search intent in their web search queries. This study introduces a K-Nearest Neighbours *(KNN, here onward)* based dual classification framework to predict search intent across two complementary taxonomies viz-a-viz Standard Search Intents *(Informational, Navigational, Commercial, Transactional)* and Health Search Intents *(Reassurance, Diagnostic, Treatment, Urgency, Courtesy, Precautionary)* and its extension to digital marketing contexts. By bringing both taxonomies in same study, the model enables a multidimensional understanding of user information-searching behaviour in the French healthcare market. These data taxonomies are generated in two sets as "Real-Human Transcripts" (20%) and "AI Simulated Text Samples" (80%) *(collectively addressed as 'Text Samples' here onward)* with purpose of addressing the possible output issues of overfitting and biasing. The process of generating both dataset samples is mentioned in 'Data Description' sections below.

These Text Samples in French were pre-processed using Natural Language Processing *(NLP, here onward)* techniques, including CamemBERT Tokenization and Torch embedding features. The KNN algorithm developed in Python was trained and evaluated on annotated datasets encompassing both standard and healthcare-related intents using Google Colab GPU A100 runtime type. Experimental results reveal that KNN achieves competitive Accuracy, AUC and F1-Scores compared to baseline classifiers, demonstrating its capacity to generalize across different intent layers. This dual-intent classification framework provides actionable insights for healthcare analytics, medical information retrieval, and patient-centred communication platforms. By simultaneously capturing generic i.e. standard search behaviour and domain-specific health concerns, the model enhances interpretability, facilitates more targeted digital health interventions, and supports the development of AI-driven decision support systems in the French healthcare sector.

The literature review of this paper examines the current state of research in search intent classification using KNN and other Deep Learning *(DL, here onward)* Models, with particular focus on healthcare applications and large-scale data processing. The review analyses 42 peer-reviewed studies spanning Machine Learning *(ML, here onward)* Algorithms, NLP in healthcare, explainable Artificial Intelligence *(AI, here onward)* and semantic similarity measures. Key findings indicate significant research gaps in French healthcare NLP, dual-intent classification taxonomies, and large-scale benchmarking frameworks. The review establishes the theoretical foundation for evaluating KNN variants against CamemBERT models using Accuracy, F1-score, and AUC metrics on both datasets containing **165, 872** text samples with **17, 773, 761** data tokens combined.

Keywords - KNN-based Search Intent Classification, Supervised Machine Learning Predictive Analytics, Semantic Similarity Analysis using KNN, Explainable AI in French Health Marketing, Search Engine Results Page *(SERP)* Navigation Behaviour Analysis

I. INTRODUCTION

France allocates ~11.4% GDP *(11.9% in 2022)* to healthcare, underscoring health importance, Alafari et al. (2025). Despite increasing volume of health queries, classifying search intent through automated systems remains a significant challenge. This classification in healthcare has emerged as a critical component, Jansen et al. (2007). To address it, our paper proposes a dual-search-intent classification framework that targets prediction of Standard Search Intents and Health Search Intents using a KNN aligned with CamemBERT model. **The straightforward yet interpretable nature of KNN, its particular utility in capturing semantic similarity, fast speed and precised detection capacity**, makes it an ideal choice for this task, where transparency and explainability are paramount, Uddin et al. (2022). Simultaneously, DL approaches, particularly BiGRU *(Bidirectional Gated Recurrent Units),* have achieved state-of-art performance in biomedical NLP tasks, Kim et al. (2023). NLP for medical search behavior in French, requires specialized handling. Prior research has demonstrated that training embeddings on clinical reports significantly enhances downstream task performance ; for instance, adapting CamemBERT on 21 million French clinical notes improved F1-scores *(a performance metric of CamemBERT)* by about 3% on clinical tasks; Basile et al. (2022). Similarly, studies on word embedding for French language in healthcare have showcased power of dense vector representations in improving concept recognition and retrieval quality; Dynomant et al. (2019) and Bazoge et al. (2023). Leveraging these insights, **our methodology** employs text preprocessing strategies *(tokenization, embedding, vectorization),* by maintaining focus on 'text samples classification' using KNN. Through predictive modeling using Accuracy, F1-Score, AUC Score, we evaluate efficacy of our dual-intent system against baseline classifier transformers. This enables a comprehensive assessment of our model's performance in capturing nuanced user intents. By integrating explainable AI in healthcare, our approach contributes to digital health solutions that are both effective and trustworthy. So, this **paper aims to:**

1. Develop a KNN classification & prediction model for dual search intent *(standard & health)* categories in French language for French healthcare marketing.
2. To evaluate the performance of KNN using multiple metrics of *Accuracy, AUC Score, F1-Score* by comparing with baseline ML *(Machine Learning)* classifiers.

II. Background:

Literature Review Methodology: This systematic and comprehensive search was conducted across multiple academic databases including PubMed, IEEE Xplore, ACM Digital Library, Springer, ResearchGate, Nature, ScienceDirect, PMC and Google Scholar by following the PRISMA guidelines. The search strategy employed combinations of keywords mentioned in "Keywords" section above. LR inclusion criteria included publications between 1951 to 2025 *(majority between 2010 to 2025)* of text classification in healthcare, marketing and AI context with particular focus on KNN and DL for Search Intents classification with explainable AI applications. Exclusion includes non-peer-review publications, studies without empirical evaluation, researches solely on image or single processing.

Search Intent Classification and Prediction in Healthcare Marketing Context: Broder (2002) established a tripartite classification of web search intents: informational, navigational, and transactional. This taxonomy was refined by Rose & Levinson (2004) to include commercial intent category. Jansen et al. (2007) demonstrated that ~80% of web queries exhibit informational intent, with navigational and transactional each comprising roughly 10%, but was unjustified for healthcare contexts. White & Horvitz (2009) identified patterns of medical concerns in search. The healthcare domain requires specialized intents accountable for varying levels of medical urgency, diagnostic uncertainty, and treatment-seeking behaviours, Amatriain (2018). This way subsequent studies have proposed healthcare-specific intent categories, but not a single research like ours. We combine standard search intents and health search intents representing a novel contribution which addresses a multifaceted health search behaviour. **Our dual-intent classification approach,** combining standard web search categories with healthcare-specific intents, represents a novel contribution addressing the multifaceted nature of health information seeking behaviour. This approach acknowledges that healthcare queries often simultaneously exhibit characteristics of multiple intent categories, requiring sophisticated classification methodologies capable of handling overlapping semantic spaces.

Limited direct studies on KNN for Healthcare Search Marketing: The intersection of KNN, search intents classification, and marketing applications represent a significant and largely unexplored research domain. Prior works in these areas such as Broder (2002), healthcare search behaviour patterns by White & Horvitz (2009), KNN applications for medical diagnosis by Uddin et al. (2022), and healthcare digital marketing frameworks of Kannan & Li (2017) reveal no comprehensive study evaluating KNN variants due to interdisciplinary complexity of combining ML algorithms with health domain knowledge and marketing applications; data privacy constraints limiting access to health search query datasets, Sbaffi & Rowley (2017), and the relatively recent emergence of sophisticated healthcare digital marketing analytics. The absence of direct precedent research, while representing a challenge for literature comparison, simultaneously positions this study as pioneering work that addresses a critical gap in academic knowledge while providing substantial practical value for healthcare organizations seeking to optimize both clinical outcomes and marketing effectiveness through intelligent search intent understanding systems, Amatriain (2018).

KNN Algorithmic Foundations and Variants: KNN algorithm, introduced by Fix & Hodges (1951), formalized by Cover & Hart (1967), operates on principle of instance-based learning, classifying data points based on majority class among KNN in feature space. Recent comprehensive analysis of **KKN variants** by Haldar et al. (2024) examined 31 KNN search methods and 12 KNN join methods, emphasizing algorithmic modifications to address the curse of dimensionality in high-dimensional medical data. **Key variants** include weighted KNN, which assigns differential importance to neighbours based on distance metrics, and adaptive KNN, which dynamically adjusts the k parameter based on local data density.

French Healthcare NLP - Current State and Challenges with Global purview: The development of NLP systems for French healthcare contexts faces unique linguistic and domain-specific challenges. Yeung et al. (2024) highlight the need for French-specific medical terminologies and automatic de-identification pipelines, emphasizing that most existing research focuses on English-language medical texts. The eHOP clinical data warehouse approach utilizes distant supervision methods to reduce manual annotation costs while achieving competitive performance on French medical text processing tasks, Azzouzi et al. (2024). Gerardin et al. (2022) demonstrates that multilingual algorithms trained on annotated clinical notes and UMLS *(Unified Medical Language System)* vocabularies, while Alonso & Contreras (2016) achieve superior performance compared to monolingual approaches; thus, represent a significant advancement for French healthcare NLP, while maintaining French-specific semantic understanding. French Clinical BERT *(Bidirectional Encoder Representations from Transformers)* and adaptations of CamemBERT *(BERT French extension)* to medical corpora have demonstrated measurable improvements in text classification and information retrieval tasks, Basile et al. (2022). Similarly, the development of French biomedical word embeddings shown improved entity recognition and semantic similarity tasks, Dynomant et al. (2019). Uddin et al. (2022) evaluated 9 different algorithmic modifications using KNN across 8 medical datasets and revealed accuracy ranges from 64.22% to 83.62%, with Hassanat KNN *(a similarity measure robust to outliers and bounded in the interval [0, 1])*. Guo et al. (2003) proposed novel KNN modifications achieving accuracy to C5.0 *(Quinlan's C5.0 algorithm)*. Recent work by Xing & Bei (2019) integration of **explainable AI** techniques with KNN demonstrated that LIME-based *(Local Interpretable Model-agnostic Explanations)* feature selection with KNN achieved

91.86% accuracy in medical health big data classification, with the explainability component identifying nine positive impact features critical for clinical decision-making.

Semantic Similarity in Healthcare Content Contexts: Pedersen et al. (2007) established biomedical **semantic similarity,** which was further lined-up for comparing knowledge-based and distributional approaches by Alonso & Contreras (2016). Garla & Brandt (2012) reported comprehensive **evaluation of semantic similarity** measures across multiple biomedical knowledge sources. Research by Sogancioglu et al. (2017) evaluated neural sentence embedding models for biomedical text, achieving Pearson correlations of 0.819 with unsupervised models and 0.871 with supervised approaches on the BIOSSES *(biomedical sentence similarity estimation system)* demonstrating superior performance over traditional methods.

Deep Learning Approaches in Healthcare NLP: Long Short-Term Memory (LSTM) networks and their bidirectional variants have proven particularly effective for capturing long-range dependencies in clinical narratives, Dernoncourt et al., 2017. Peng et al. (2019) evaluated BERT and ELMo *(Embeddings from Language Models)* models across ten biomedical datasets, establishing transfer learning baselines for the medical domain; their pre-training significantly improves performance on medical text classification tasks. However, data annotation remains a significant bottleneck in clinical NLP, Spasic & Nenadic (2020), with most studies limited to small datasets potentially limiting model generalizability. Here our model promises a successful big data size testing.

Large-Scale Data Processing and Scalability Challenges in Healthcare NLP: The processing of large-scale healthcare datasets presents unique computational and methodological challenges. Most academic studies in healthcare NLP operate on relatively small datasets (typically <10,000 samples) due to annotation costs and privacy constraints Spasic & Nenadic (2020). Our research utilizes AI simulation methodology to generate $165,872$ samples with $17,773,761$ data tokens, representing a 10-50x increase in scale. The application of dimensionality reduction techniques becomes critical for **large-scale healthcare NLP applications.** Truncated Singular Value Decomposition (SVD) has proven effective for reducing TF-IDF *(Term Frequency - Inverse Document Frequency)* feature spaces while preserving semantic relationships relevant for classification tasks. The integration of information gain and principal component analysis approaches has shown particular promise for healthcare text classification applications, Xing et Bei (2019).

Explainable AI in Healthcare Applications: The deployment of AI systems in healthcare contexts necessitates exceptional attention to explainability and interpretability due to the high-stakes nature of medical decision-making. Amann et al. (2020) emphasize that explainability represents a fundamental requirement for healthcare AI systems, as opaque algorithms pose threats to core ethical values in medicine. Explainable AI integration with KNN in healthcare reveals six distinct methodological categories: feature-oriented methods, global methods, concept models, surrogate models, local pixel-based methods, and human-centric approaches, Sadeghi et al. (2024). Zhang et al. (2022) demonstrate that LIME-based explanations can effectively identify critical features influencing classification decisions while maintaining model performance. Research by Pietila & Moreno-Sanchez (2023) introduces Clinical Explainability Failure (CEF) and Explainability Failure Ratio (EFR) metrics, addressing cases where classification accuracy is high but explanations fail to meet clinical requirements. The Modality-Specific Feature Importance (MSFI) metric addresses clinical requirements for understanding AI decision processes across different input modalities, ensuring explanations meet practical clinical needs.

Research Gaps Analysis: While prior work as reviewed here has successfully leveraged transformer-based models for healthcare text classification, and some exploratory studies have applied KNN to biomedical similarity tasks, but there is no documented benchmarking of KNN for intent prediction in the French or European healthcare systems; and integration of dual-intent taxonomies *(like ours)* remains absent till date. So, we intend to address following research gaps:

1. **Scalable French Healthcare NLP Model**: To fulfil gap of limited evaluation beyond 10,000 by using 165, 872 French Text Samples with head-to-head comparison of KNN and DL, also going beyond English orbits.
2. **Dual-Intent-categories classification hybrid approach**: Our paper integrates KNN and DL methodologies for hybrid efficiency and performance in this regard.
3. **Domain-specific Real-time Evaluation:** Our study advances in healthcare AI evaluation with crucial ingredients to develop efficient real-time algorithms.

Also, following promising research directions emerge from this literature review:

1. **Hybrid Approaches**: Integration of KNN and DL methodologies for combined efficiency and performance, which our paper aims to address.
2. **Multilingual Transfer Learning**: Extension of French Healthcare NLP through cross-lingual knowledge transfer going beyond English orbits.
3. **Real-Time Processing**: Development of efficient algorithms for real-time healthcare search intent classification as the prime invention of our study.

III. Research Methodology:

Dataset Description: We employed **128,847 Text Samples** to train our KNN model for detection of **standard search intents**, and **37,024 Text Samples** to detect **Health Search Intents** respectively**.** Each dataset is composed of two main components:

Real-human data samples (20% of total data):

- Text samples extracted from 627 video tests transcripts - 10%. *For these 161 participants 99% of the age from 20 to 26 years from Northern France Market were recorded in videos for their Search Results Navigation Behaviour on various Search Engines, 4 test videos per participant for 4 different search scenarios given. Then these videos were transcribed into text, then these transcripts were fragmented into 9335 sensible text samples, combined making 815340 data tokens.*
- Text samples collected from YouTube texts - 10%. These text samples comprised of Healthcare sector comments and remarks on health-related videos in French language.

Synthetic yet realistic LLMs AI Simulation data samples (80% of total data):

- Texts samples generated with **Chat-GPT** and **Mistral** models. These samples were then downloaded on local machine using LLMs API Keys anonymously. Appendix-1 below describes our "Query Text Generation Prompts" and "YouTube API".
- Generation process inspired by authentic transcripts created through contextual *(standard and health intents)* search prompt instructions given to GPTs, followed by systematic phases of **diversification, validation, and filtering**, designed to both optimize the quality and correct potential inconsistencies in outputs.

This triple-source approach adjusted into two datasets ensures the **representativeness** of real-world discourse and the **diversity** required for robust model training. Figure-1 below demonstrates this treatment:

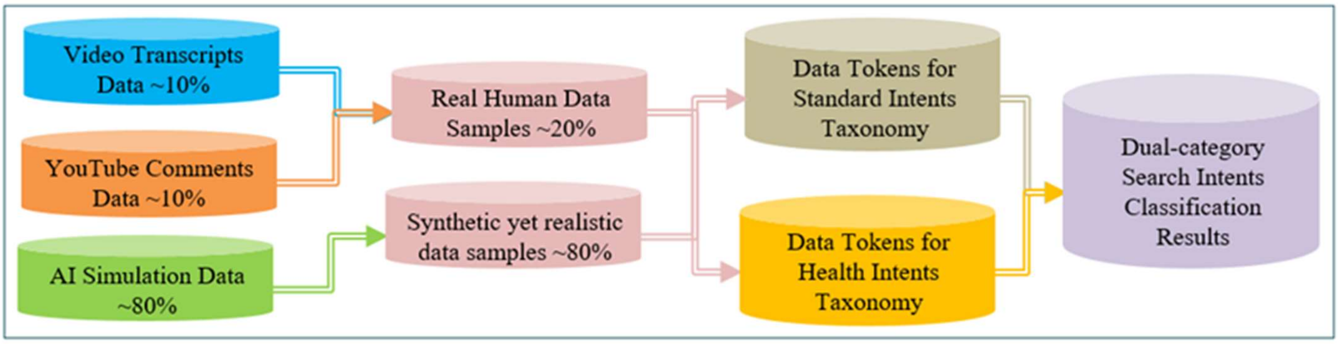
Video Transcripts Data ~10%
YouTube Comments Data ~10%
AI Simulation Data ~80%
Real Human Data Samples ~20%
Synthetic yet realistic data samples ~80%
Data Tokens for Standard Intents Taxonomy
Data Tokens for Health Intents Taxonomy
Dual-category Search Intents Classification Results

Figure-1: Data Creation and Treatment Pipeline

Why mix of Real Data with LLMs AI Simulated Data: This approach was used to ensure the diversity of data samples for better quality, to give our model more efficiency and to attain the minimum data requirements of the model. Without AI-simulation data it was proving difficult to handle output overfitting issue and the overrated results biasing. Problem of prediction quality due to limited data size *(just real transcripts data)* was another challenge to handle. So, to improve the quantity with good quality, we generated new AI data samples and mixed with the real data.

Labelling approach: After generation of the data, it needed an efficient labelling. To efficiently label the data, we opted for an automated approach, as the manual process would be too time-consuming given the large volume of data to process. We used GPT *(Generative Pre-trained Transformer)* to perform the labelling, providing clear definitions for each intent, strict annotation guidelines, and a required output format in the form of a multi-dimensional binary vector - one binary value per intent. For each text, a binary label is assigned to every intent: 1 if the intent is detected according to the intent definition and the labelling rules, and 0 if not. Appendix-2 below presents the sample of Prompts and Rules we used for Data Labelling.

Intent definitions and general rules for Labelling: We used dual categories as "Standard Search Intents" and "Health Search Intents". Here below is the definition of each intent with its category:

Standard Search Intents Category:

- **Navigational Intent:** The user wants to access online a specific site, service, or page they already know.
- **Informational Intent:** The user seeks to acquire knowledge, learn, explore, get informed or understand something online, without necessarily wanting to act afterward.
- **Transactional Intent:** The user wants to carry out a specific action, such as purchasing, subscribing, downloading, or booking. This reflects behaviour oriented toward an immediate goal.
- **Commercial Intent:** The user is conducting pre-purchase research: comparing, checking reviews, or looking for alternatives. They are not yet ready to buy, but are seriously considering their options.

Health Search Intents Category:

- **Precautionary Intent:** Seeking information to prevent health issues or maintain well-being.
- **Reassurance Intent:** Seeking confirmation or peace of mind regarding one's health.
- **Diagnostic Intent:** Identifying the cause of symptoms or health problems.
- **Treatment Intent:** Finding solutions or remedies for an already diagnosed condition.
- **Courtesy Intent:** Searching for health information on behalf of someone else *(care, concern, or support).*
- **Urgency Intent:** Addressing immediate health problems that require quick action.

The presence of each intent is evaluated independently by assigning **label 1 for presence and 0 for absence**. The following rules were applied:

- An intent only receives **1** if it is clearly and explicitly expressed, with concrete actions or associated words/phrases.
- Assumptions or interpretations from weak or vague signals are not considered, and inference is limited to what is literally stated or strongly implied. If an intent is not present, its score is **0**.
- If the signal is vague, missing, or purely hypothetical, assign the value **0 (absent)**.
- A sentence may contain multiple intents or at least one intent.
- Each intent is evaluated independently.

Further in process to ensure the quality and better understand the nature of our data, its token analysis and similarity analysis were done as explained below.

Data Token Analysis Reports for all types of Data used:

Data Tokens for Standard Intent Taxonomy combined of both datasets:

- Total number of text samples: **128,847**
- Total number of data tokens of text samples: **14,493,837**
- Descriptive Statistics of Token Counts per Text is presented in Table-1 below:

Statistic	Value
count	128847.00

mean	112.49
std	41.61
min	5.00
25%	85.00
50%	120.00
75%	143.00
max	1244.00

Table-1: Descriptive Statistics of Token Count for Standard Search Intent Category

Data Tokens for Health Intent Taxonomy combined of both datasets:

- Total number of text samples: **37025**
- Total number of data tokens of text samples: **3,279,924**
- Descriptive Statistics of Token Counts per Text is presented in Table-2 below:

Statistic	Value
count	37021.00
mean	88.60
std	148.30
min	8.00
25%	44.00
50%	58.00
75%	83.00
max	6433.00

Table-2: Descriptive Statistics of Token Count for Health Search Intent Category

Combined Total of Text Samples and Data Tokens:

- Total number of text samples: **165, 872**
- Total number of data tokens of text samples: **17, 773, 761**

Text Samples Similarity Analysis:

Result of similarity analysis of standard intent data: Table-3 below shows the global similarity statistics calculated on 100% of Text Samples *(128,847),* based on a random sampling of 120,000 pairs *(TF-IDF across the entire corpus, optimized sparse computations)*:

Statistic	Value
Number of Text Samples	128847
Number of pairs evaluated	120000
Minimum	0.000
Q1 (25%)	0.048
Median (50%)	0.083
Q3 (75%)	0.123
Maximum	0.981
Mean	0.130

Table-3: Descriptive Statistics of Token Count for Health Search Intent Category

Interpretation: The distribution curve of these similarities shown in Figure-2 highlights followings:

- A strong concentration between **0.05 and 0.15** → corpus is globally diverse;
- A right tail with some very close pairs (up to ~0.98) → likely presence of nearly identical reformulations.
- We notify very low presence of nearly identical texts samples.

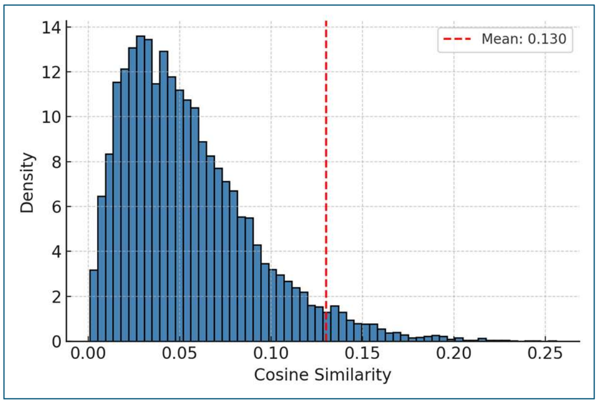

Figure-2: Distribution of cosine similarities for Standard Intent Category

Result of similarity analysis of health-care intention data: Table-4 below shows the result of the similarity study of the second dataset *(37,024 text samples),* based on a random sampling of 80,000 pairs *(TF-IDF across the entire corpus, optimized sparse computations)*:

Statistic	Value
Number of Text Samples	37,024
Number of pairs evaluated	~80,000
Minimum	0.000

Q1 (25%)	0.022
Median (50%)	0.050
Q3 (75%)	0.139
Maximum	0.98
Mean	0.118

Table-4: Descriptive Statistics of Token Count for Health Search Intent Category

Interpretation: The distribution curve of these similarities shown in Figure-3 highlights followings:

- The distribution is centered on low values (0.02 – 0.14), so the corpus is globally diverse.
- The mean of ~0.12 is very close to that observed, which confirms the general trend.
- The presence of pairs at 0.98 indicates the existence of nearly identical text samples.
- We notify very low presence of nearly identical texts samples.

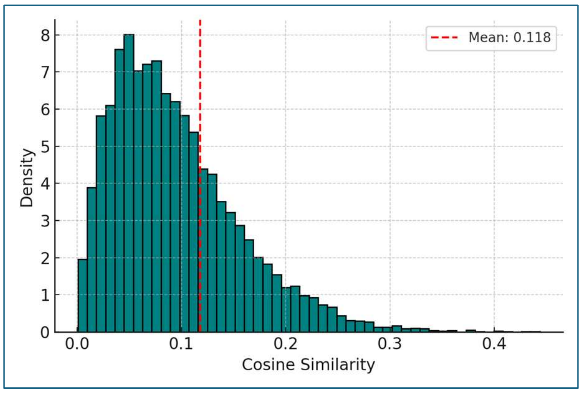

Figure-3: Distribution of cosine similarities for Health Intent Category

Percentage Intents Distribution of both categories: Continuing in pipeline of describing our data, this distribution was calculated considering the presence of an intent percentage against 1. Here to note that the Text Samples express at least one intent and can express multiple intents (multi-labels) same time.

Distribution of standard intents in the data
Commercial: 10.00% · Informational: 84.00% · Navigational: 20.50% · Transactional: 23.44%

Distribution of healthcare intents in the data
Precautionary: 14.20% · Reassurance: 13.00% · Diagnostic: 30.33% · Treatment: 54.36% · Urgency: 22.15% · Courtesy: 14.50%

Supervised Machine Learning Model Protocol: The data generated and prepared through above protocol was then treated through Google Colab GPU A100 on KNN and CamemBERT Multi-Label Text classification to calculate output metrics as explained in Figure-3. This set was tested for both Standard and Health Search Intents dual categories.

Standard Performance Metrics: For the evaluation of text samples classification we used Accuracy, F1-score, and Area Under the ROC (Receiver Operating Characteristic) Curve (AUC) metrics. As, the selection of appropriate metrics depends on specific application requirements and class distribution characteristics. F1-score has emerged as particularly valuable for healthcare applications due to its balanced consideration of precision and recall, addressing class imbalance issues, Wieland-Jorna et al. (2024) common in health texts like our datasets, the weighted F1-score variant provides additional robustness for multi-class scenarios with uneven class distributions. The Area Under the ROC Curve (AUC) has been extensively validated as a robust metric for binary classification performance, with foundational work by Hanely & McNeil (1982) and Bradely (1997) establishing its theoretical foundations, while Fawcett (2006) provided comprehensive guidance on ROC analysis interpretation and application. For text classification and NLP applications, Sebastiani (2002) and Yang & Liu (1999) established that metric selection should align with specific task requirements and dataset characteristics, while recent comparative analyses by Davis & Goadrich (2006) and Chicco & Jurman (2020) provide frameworks for selecting appropriate metrics based on dataset properties, class distribution, and application requirements, supporting the use of multiple complementary metrics including accuracy, AUC, and F1-scores for comprehensive model evaluation in complex domains like healthcare search intent classification.

Cross-validation for Overfitting and Generalizability Assessment: Cross-validation strategies become critical for ensuring model generalizability, particularly given the limited size of most healthcare NLP datasets. Systematic review reveals that most studies employ train/test splits and k-fold cross-validation, though many fails to adequately address overfitting concerns (Hossain et al., 2023). The present research addresses generalizability concerns through comprehensive evaluation across multiple data sizes *(1,000 to 120,000 samples),* enabling assessment of model performance scaling characteristics and identification of optimal deployment scenarios.

Research Model Pipeline: In accordance to the Processing Workflow presented in Figure-4, below is a sequential execution detailing of our Supervised Machine Learning Model:

Step 1: Environment Setup and Dependencies Configurations: Installing required libraries of Pandas, Numpy, sklearn train_test_split model, sklearn accuracy score, f1 score, roc auc score metrics, sklearn onevsrestclassifier multiclass, sklearn kneighboursclassifier, CamemBERT tokenizer and CamemBERT model transformers, torch, tqdm and google colab files; importing essential modules for data processing, model training, and evaluation, Configuring Google Collab GPU A100 runtime type for optimal computational. Google Colab GPU A100 delivers up to ~156 TFLOPS *([TeraFLoating point Operations Per Second] - TF32 AI precision)* and has 40 GB VRAM *(Video Random Access Memory)* with >1.5 TB/s bandwidth, making it one of the fastest GPUs available for machine learning research. It's ideal for training/fine-tuning very large neural networks and for running heavy inference workloads without memory bottlenecks.

Step 2: ML Model Initialization and Data Ingestion: Defining target label columns for dual intent categories as ['Commercial', 'Informational', 'Navigational', 'Transactional'] and ['Reassurance', 'Diagnostic', 'Treatment', 'Precautionary', ‘Urgency’, ‘Courtesy’]; setting model to evaluation mode *(no fine-tuning of transformer weights);* loading model onto available device *(CUDA if available);* Uploading labelled Text Samples data in Excel using pandas with robust handling of missing values.

Step 3: Data Pre-processing: Initializing CamemBERT tokenizer and pre-trained model *(camemBERTt-base);* binarizing multi-label target variables *(convert to 0/1 binary classification).*

Step 4: Data Features Engineering and Partitioning: Extraction via Transformer Embeddings such as extracting text features from 'File Text' column with null value imputation; implementing mean pooling strategy for sequence-level representation; batch processing of text samples data through CamemBERT tokenizer *(max_length=256),* generating 768-dimensional dense embeddings using pre-trained transformer; applying attention mask weighting for variable-length sequences and storing embeddings as fixed-size numerical feature vectors. After Data Engineering, the engineered data was performed on stratified train-test split *(80/20 ratio, random_state=42)* while ensuring balanced representation across all label classes and maintaining data integrity through consistent indexing.

Step 5: Model Architecture Definition and Training: Configuring of k-Nearest Neighbors classifier then implementing One-vs-Rest strategy for multi-label classification; applying distance weighting to neighbor contributions, using brute force algorithm for cosine distance computation. And at the end training classifier on embedded feature representations.

Step 6: Model Inference and Prediction: Generating binary predictions for test set, extracting class probabilities using robust probability estimation; and handling sklearn version compatibility for OneVsRestClassifier probability output.

Step 7: Model Evaluation, Metrics and Results Export: Calculating three performance metrics per label class to report comprehensive performance statistics:

- **Accuracy:** Proportion of correct predictions
- **F1-Score:** Harmonic mean of precision and recall
- **AUC-ROC:** Area under receiver operating characteristic curve

Creating comparison DataFrame with true vs predicted labels; exporting results to Excel format for further analysis.

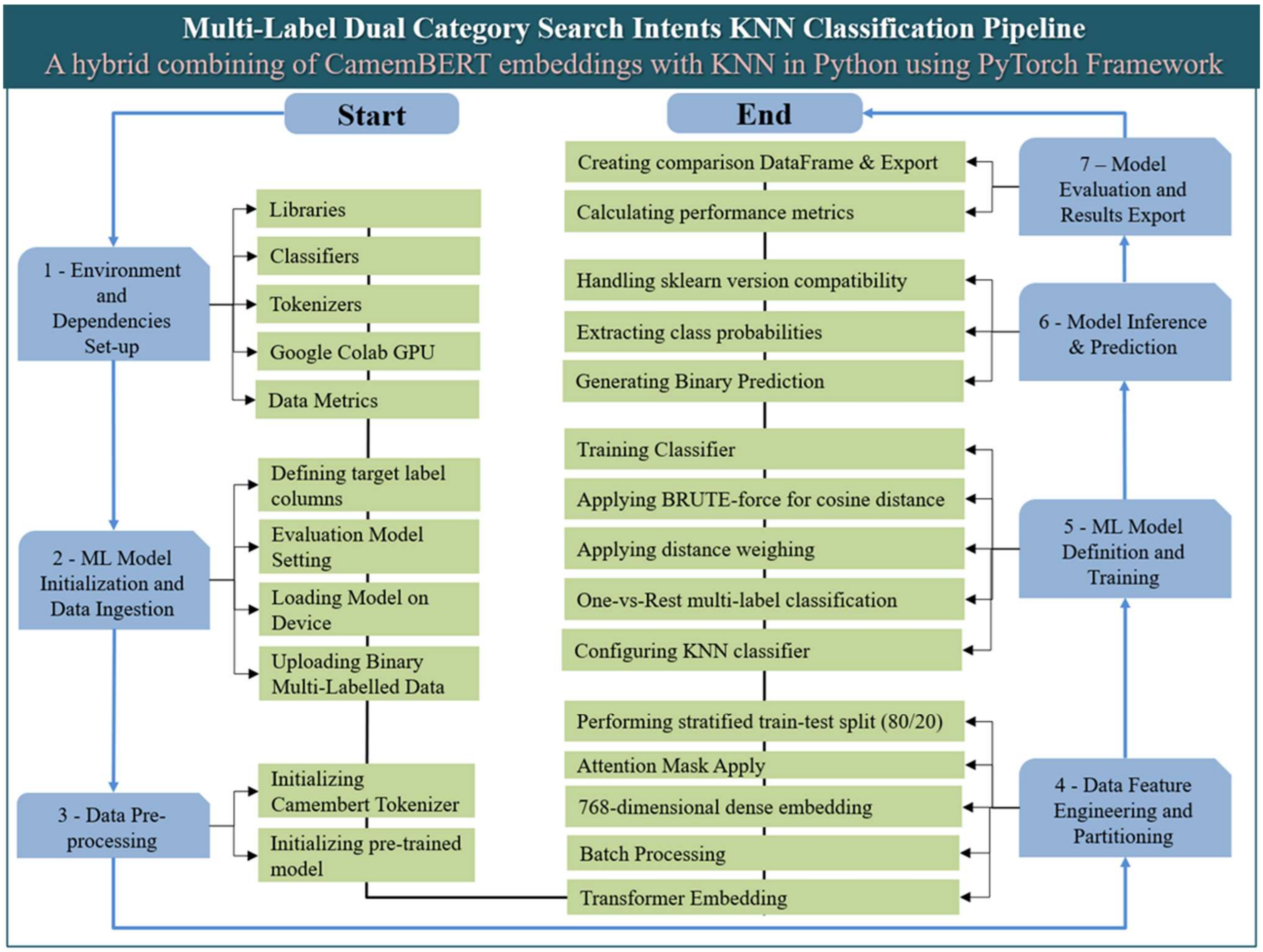
Multi-Label Dual Category Search Intents KNN Classification Pipeline
A hybrid combining of CamemBERT embeddings with KNN in Python using PyTorch Framework
Start
End
1 - Environment and Dependencies Set-up
Libraries
Classifiers
Tokenizers
Google Colab GPU
Data Metrics
2 - ML Model Initialization and Data Ingestion
Defining target label columns
Evaluation Model Setting
Loading Model on Device
Uploading Binary Multi-Labelled Data
3 - Data Pre-processing
Initializing Camembert Tokenizer
Initializing pre-trained model
4 - Data Feature Engineering and Partitioning
Performing stratified train-test split (80/20)
Attention Mask Apply
768-dimensional dense embedding
Batch Processing
Transformer Embedding
5 - ML Model Definition and Training
Training Classifier
Applying BRUTE-force for cosine distance
Applying distance weighing
One-vs-Rest multi-label classification
Configuring KNN classifier
6 - Model Inference & Prediction
Handling sklearn version compatibility
Extracting class probabilities
Generating Binary Prediction
7 – Model Evaluation and Results Export
Creating comparison DataFrame & Export
Calculating performance metrics

Figure-4: Search Intents Classification KNN Pipeline

Data Metrics Calculations and Results Analysis:

Table-5 shows our Training Performance Results for Standard Search Intents:

Intent Type	Accuracy	F1 Score	AUC
Commercial	0.9825	0.9071	0.9947
Informational	0.9744	0.9849	0.9931
Navigational	0.9779	0.9445	0.9946
Transactional	0.9735	0.9415	0.9936

Table-5

Table-6 shows our Training Performance Results for Health Search Intents:

Intent Type	Accuracy	F1 Score	AUC
Precautionary	0.9525	0.8009	0.9681
Diagnostic	0.9313	0.8817	0.9799
Treatment	0.9136	0.9210	0.9676
Courtesy	0.9602	0.8355	0.9683
Urgency	0.9783	0.9477	0.9845
Reassurance	0.9780	0.9104	0.9820

Table-6

Interpretation of above Numerical Results: Our model reveals several scientifically relevant insights when contextualized through corpus similarity analysis, intent distribution, and the semantics of applied metrics.

First, the similarity analysis confirms that our datasets *(standard intents and health intents)* are globally diverse, with median cosine similarity values between 0.05-0.08; which implies that the model was exposed to heterogeneous linguistic formulations, thereby reducing the risk of overfitting to repetitive patterns and enhancing generalizability. The presence of a right-tail distribution with high similarity pairs *(up to 0.98)* indicates a limited set of duplicates or near-duplicates, which is expected in large-scale corpora and does not compromise overall diversity; and that is particularly critical for intent detection, where subtle lexical variations often signal different underlying user motivations.

Second, the distribution of intents across datasets shows significant imbalance, for example informational intent dominates standard taxonomy *(84%),* while treatment intent accounts for more than half of healthcare dataset *(54.36%).* Such imbalance introduces classification challenges, as models trained on skewed distributions may over-predict majority classes. This makes F1-score a more reliable indicator than accuracy. **Accuracy score,** which measures the proportion of correctly classified instances, remains very high across all intents *(0.91–0.98),* but by itself could mask poor recognition of minority categories. **F1-score**, by harmonizing precision and recall, directly addresses this issue and thus provides a more nuanced view. The results show consistently strong F1-scores, with values ranging between 0.80 to 0.98 while above 0.90 for most categories except three, Precautionary *(0.80),* Diagnostic *(0.88)* and Courtesy *(0.83);* but still robust performance for these three as well. These scores confirm that the model is not merely biased toward majority intents, but is capable of balancing sensitivity and specificity even under class imbalance conditions.

Third, the **AUC values** *(0.96–0.99)* further corroborate discriminative power of our model that demonstrates that the classifier maintains strong ranking ability, i.e., it can consistently separate positive from negative examples even when thresholding is varied. This metric is particularly important in healthcare contexts where the costs of false positives and false negatives differ depending on intent type *(e.g., urgency vs reassurance).* The near-perfect AUC scores suggest the classifier captures meaningful semantic boundaries between intents rather than relying on incidental correlations.

Taken together, the combination of high accuracy, strong F1-scores across balanced and imbalanced classes, and near-optimal AUC demonstrate that the KNN approach, supported by CamemBERT embeddings, is highly effective for dual-intent classification. Importantly, the corpus diversity revealed by similarity analysis strengthens the validity of these results, indicating that the model's predictive quality is not an artefact of dataset redundancy but rather a reflection of genuine semantic learning.

IV. Discussion and Contribution

The integration of large-scale AI simulation data with real-human data represents a significant methodological advancement, enabling large-scale evaluation in a dual intent classification framework. **Our work contributes to fulfil** following gaps:

1. **Practical Relevance:** Our work provides strategic and actionable insights for digital health platforms in France, improving medical information retrieval and search query handling; which all in combination should be of high interest for marketing managers and strategists.
2. **Novelty Contribution:** Our model demonstrates how a lightweight, interpretable model can achieve competitive results compared to heavier DL models. As the gaps and limits expressed by Azzouzi et al. (2024) and Gerardin et al. (2022) and single taxonomies limit cited by Jansen et al. (2007); our paper presents a hybrid ML model with three sources taxonomies treated for dual-intents-classification with 100% French health search data *(both real and synthetic sets).*
3. **Benchmarking:** Outputs of our Model ranges between 0.9825 to 0.9136 for accuracy score, 0.8009 to 0.9849 for F1 score and 0.9681 to 0.9947 for AUC score; which are in all cases better than KNN performances mentioned by Uddin et al. (2022) and *Sogancioglu et al. (2017).*
4. **Explainable AI Integration:** Zhang et al. (2022) cited the demand of explainable AI solutions in healthcare. Our Model offers inherent interpretability advantages with improved medical information retrieval and search quality handling over black-box models.

V. Limitation & Future Scope:

Improving prediction by adding sensibility study is yet to test. The values and method chosen by us are good enough but with more resources and time, we want to optimize our results further. Though problematic but by injecting and testing with more real data can further improve outputs. We tried testing of our model with only real data as well and got appreciable results with two of our three metrics; but still were not up to our satisfaction, so it requires to further supplementary tests; which we have planned to try with another better Machine Learning Framework.

VI. Conclusion:

Our model achieves a high level of predictive quality, capable of distinguishing both standard and health search intents with robustness and generalizability. Its balanced performance across metrics highlights its scientific relevance and practical utility for healthcare information retrieval and digital health marketing applications. It also establishes a large-scale methodological benchmark for explainable AI in French-language healthcare NLP. Our research and its analysis bridges significant research gaps in French healthcare NLP, large-scale benchmarking frameworks, and especially the dual-intent classification methodologies. The historical identified performance ranges *(64.22-83.62% for KNN variants Uddin et al. (2022), 0.819-0.871 Pearson correlation for neural approaches, Sogancioglu et al. (2017))* while ours output ranges between 0.9825 to 0.9136 for accuracy score, 0.8009 to 0.9849 for F1 score and 0.9681 to 0.9947 for AUC score; which promises a robust modelling.

REFERENCES in APA Format:

1. Alafari, F., Driss, M., & Cherif, A. (2025). Advances in natural language processing for healthcare: A comprehensive review of techniques, applications, and future directions. *Computer Science Review*, *56*, 100725. https://www.sciencedirect.com/science/article/abs/pii/S1574013725000024?
2. Basile Dura, C. J., Tannier, X., Calliger, A., Bey, R., Neuraz, A., & Flicoteaux, R. (2022). Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. *arXiv preprint arXiv:2207.12940*. https://arxiv.org/abs/2207.12940?
3. Dynomant, E., Lelong, R., Dahamna, B., Massonnaud, C., Kerdelhué, G., Grosjean, J., ... & Darmoni, S. J. *Word embedding for the French natural language in health care: comparative study. JMIR Med Inform. 2019 Jul 29; 7 (3): e12310. 10.2196/12310.* https://medinform.jmir.org/2019/3/e12310/PDF
4. Bazoge, A., Morin, E., Daille, B., & Gourraud, P. A. (2023). Applying natural language processing to textual data from clinical data warehouses: systematic review. *JMIR medical informatics*, *11*, e42477. https://medinform.jmir.org/2023/1/e42477/PDF
5. Névéol, A., Dalianis, H., Velupillai, S., Savova, G., & Zweigenbaum, P. (2018). Clinical natural language processing in languages other than English: opportunities and challenges. *Journal of biomedical semantics*, *9*(1), 12. https://jbiomedsem.biomedcentral.com/articles/10.1186/s13326-018-0179-8?
6. Azzouzi, M. E., Coatrieux, G., Bellafqira, R., Delamarre, D., Riou, C., Oubenali, N., ... & Bouzillé, G. (2024). Automatic de-identification of French electronic health records: a cost-effective approach exploiting distant supervision and deep learning models. *BMC Medical Informatics and Decision Making*, *24*(1), 54. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02422-5
7. Gérardin, C., Wajsbürt, P., Vaillant, P., Bellamine, A., Carrat, F., & Tannier, X. (2022). Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, *128*, 102311. https://dl.acm.org/doi/abs/10.1016/j.artmed.2022.102311
8. Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256. https://www.nature.com/articles/s41598-022-10358-x
9. Zhang, Y., Weng, Y., & Lund, J. (2022). Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, *12*(2), 237. https://pmc.ncbi.nlm.nih.gov/articles/PMC8870992/
10. Jansen, B. J., Booth, D. L., & Spink, A. (2007, May). Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1149-1150). https://www.researchgate.net/publication/221023370_Determining_the_user_intent_of_web_search_engine_queries
11. Garla, V. N., & Brandt, C. (2012). Semantic similarity in the biomedical domain: an evaluation across knowledge sources. *BMC bioinformatics*, *13*(1), 261. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-261
12. Kim, H., Lee, J., Moon, S., Kim, S., Kim, T., Jin, S. W., ... & Park, J. R. (2023). Visual field prediction using a deep bidirectional gated recurrent unit network model. *Scientific Reports*, *13*(1), 11154. https://www.nature.com/articles/s41598-023-37360-1
13. Broder, A. (2002, September). A taxonomy of web search. In *ACM Sigir forum* (Vol. 36, No. 2, pp. 3-10). New York, NY, USA: ACM. https://dl.acm.org/doi/10.1145/792550.792552

14. Rose, D. E., & Levinson, D. (2004, May). Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web* (pp. 13-19). https://www.researchgate.net/publication/2947641_Understanding_User_Goals_in_Web_Search
15. Amatriain, X. (2018). NLP & healthcare: Understanding the language of medicine. *Medium*. https://medium.com/curai-tech/nlp-healthcare-understanding-the-language-of-medicine-e9917bbf49e7
16. White, R. W., & Horvitz, E. (2009). Cyberchondria: studies of the escalation of medical concerns in web search. *ACM Transactions on Information Systems (TOIS)*, *27*(4), 1-37. http://ryenwhite.com/papers/WhiteTOIS2009.pdf
17. Fix, E., & Hodges, J. L. (1951). Discriminatory analysis, nonparametric discrimination. https://www.jstor.org/stable/1403797
18. Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27. https://ieeexplore-ieee-org.ressources-electroniques.univ-lille.fr/stamp/stamp.jsp?tp=&arnumber=1053964
19. Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., & Khraisat, A. (2024). Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data*, *11*(1), 113. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00973-y
20. Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003, November). KNN model-based approach in classification. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"* (pp. 986-996). Berlin, Heidelberg: Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-540-39964-3_62
21. Xing, W., & Bei, Y. (2019). Medical health big data classification based on KNN classification algorithm. *Ieee Access*, *8*, 28808-28819. https://www.researchgate.net/publication/337505760_Medical_Health_Big_Data_Classification_Based_on_KNN_Classification_Algorithm
22. Dernoncourt, F., Lee, J. Y., Uzuner, O., & Szolovits, P. (2017). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, *24*(3), 596-606. https://pubmed.ncbi.nlm.nih.gov/28040687/
23. Pedersen, T., Pakhomov, S. V., Patwardhan, S., & Chute, C. G. (2007). Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, *40*(3), 288-299. https://pubmed.ncbi.nlm.nih.gov/16875881/
24. Peng, Y., Yan, S., & Lu, Z. (2019). Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*. https://aclanthology.org/W19-5006/
25. Spasic, I., & Nenadic, G. (2020). Clinical text data in machine learning: systematic review. *JMIR medical informatics*, *8*(3), e17984. https://pmc.ncbi.nlm.nih.gov/articles/PMC7157505/
26. Alonso, I., & Contreras, D. (2016). Evaluation of semantic similarity metrics applied to the automatic retrieval of medical documents: An UMLS approach. *Expert Systems with Applications*, *44*, 386-399. https://repositorio.comillas.edu/xmlui/handle/11531/15474
27. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*(1), 310. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01332-6
28. Au Yeung, J., Shek, A., Searle, T., Kraljevic, Z., Dinu, V., Ratas, M., ... & Teo, J. T. (2024). Natural language processing data services for healthcare providers. *BMC medical informatics and decision making*, *24*(1), 356.https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02713-x
29. Soğancıoğlu, G., Öztürk, H., & Özgür, A. (2017). BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, *33*(14), i49-i58. https://pubmed.ncbi.nlm.nih.gov/28881973/
30. Wieland-Jorna, Y., van Kooten, D., Verheij, R. A., de Man, Y., Francke, A. L., & Oosterveld-Vlug, M. G. (2024). Natural language processing systems for extracting information from electronic health records about activities of daily living. A systematic review. *JAMIA open*, *7*(2), ooae044. https://academic.oup.com/jamiaopen/article/7/2/ooae044/7681769
31. Kannan, P. K., Hongshuang “Alice” Li. (2017). Digital marketing: A framework, review and research agenda. *International journal of research in marketing*, *34*(1), 22-45. https://www.sciencedirect.com/science/article/abs/pii/S0167811616301550
32. Sbaffi, L., & Rowley, J. (2017). Trust and credibility in web-based health information: a review and agenda for future research. *Journal of medical Internet research*, *19*(6), e218. https://www.jmir.org/2017/6/e218/
33. Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145-1159. https://www.sciencedirect.com/science/article/abs/pii/S0031320396001422
34. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36. https://pubs.rsna.org/doi/10.1148/radiology.143.1.7063747
35. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8), 861-874. https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X
36. Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), 1-47. https://dl.acm.org/doi/10.1145/505282.505283
37. Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and*

development in information retrieval (pp. 42-49). https://dl.acm.org/doi/10.1145/312624.312647

38. Davis, J., & Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233-240). https://dl.acm.org/doi/10.1145/1143844.1143874
39. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 6. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6413-7
40. Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in biology and medicine*, *155*, 106649. https://www.scienceDirect.com/science/article/abs/pii/S0010482523001142
41. Pietilä, E., & Moreno-Sánchez, P. A. (2023, September). When an explanation is not enough: an overview of evaluation metrics of explainable AI systems in the healthcare domain. In *Mediterranean Conference on Medical and Biological Engineering and Computing* (pp. 573-584). Cham: Springer Nature Switzerland. https://www.researchgate.net/publication/377107192_When_an_Explanation_is_not_Enough_An_Overview_of_Evaluation_Metrics_of_Explainable_AI_Systems_in_the_Healthcare_Domain
42. Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., ... & Pardalos, P. M. (2024). A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, *118*, 109370. https://www.sciencedirect.com/science/article/pii/S0045790624002982

Appendix-1: Query Text Generation Prompts and YouTube API

The prompts below were used in several attempts to promote diversity in the generated texts through G-AI Simulation using their APIs; and YouTube API. G-AI tools used include Chat-GPT, Mistral, DeepSeek and Perplexity.

Les invites ci-dessous ont été utilisées dans plusieurs tentatives pour promouvoir la diversité dans les textes générés via G-AI Simulation en utilisant leurs API; et l'API YouTube. Les outils G-AI utilisés incluent Chat-GPT, Mistral, DeepSeek et Perplexity.

Prompt de génération de textes pour les intentions standards: Générer 1000 textes en s'inspirant des textes dans le excel joint. il faut des textes suffisamment différents (vocabulaires, expressions, liaisons logiques, etc). La similarité des textes générés doit varier entre 5 et 80% avec une similarité moyenne proche de 20%. Les textes doivent exprimer au moins une intention parmis les intentions suivantes:

- **Intention transactionnelle:** L'utilisateur est prêt à effectuer une action spécifique, comme acheter, s'inscrire, télécharger ou réserver.
  Cela reflète un comportement orienté vers un objectif immédiat.
- **Intention navigationnelle:** L'utilisateur veut accéder à un site, service ou page spécifique qu'il connaît déjà. Le moteur de recherche est utilisé comme un raccourci.
- **Intention informationnelle:** L'utilisateur cherche à acquérir des connaissances ou à comprendre quelque chose, sans forcément vouloir agir ensuite. L'objectif est d'apprendre, d'explorer ou de s'informer.
- **Intention commerciale:** L'utilisateur effectue des recherches avant achat : il compare, consulte des avis, ou cherche des alternatives. Il ne souhaite pas encore acheter, mais il considère sérieusement ses options.

Prompt de génération de textes pour les intentions de type médicale « health care intentions »: Générer 1000 textes en s'inspirant des textes dans le excel joint. il faut des textes suffisamment différents (vocabulaires, expressions, liaisons logiques, etc). La similarité des textes générés doit varier entre 5 et 80% avec une similarité moyenne proche de 20%. Les textes doivent exprimer au moins une intention parmis les intentions suivantes:

- **Intention de precaution**: Rechercher des informations pour prévenir les problèmes de santé ou préserver son bien-être.
- **Intention diagnostique**: Identifier la cause de symptômes ou de problèmes de santé.
- **Intention thérapeutique:** Trouver des solutions ou des remèdes à une affection diagnostiquée.
- **Intention de courtoisie:** Rechercher des informations de santé pour autrui (soin, inquiétude, accompagnement).
- **Intention d'urgence:** Problèmes de santé immédiats nécessitant une action rapide.
- **Intention de se rassurer**: Rechercher une confirmation ou une tranquillité d'esprit concernant sa santé.

Requêtes sur l'API Youtube: Exemples de Môts clés utilisés pour les requêtes sur Youtube: « Allo docteur », « appel d'urgence + secteur médicale », « consultation + santé », « service d'urgence + accident », « SAMU », « Produits pharmaceutiques », et d'autres môts clés.

Les résultats des requetes sont manuellement filrés puis transcrit avec Whisper, une IA de OpenAI. A la suite de la transcription, les transcriptes sont découpés en textes ayant un sens sémantique clair et au moins une intentions parmi les intentions étudiées. La longueur des textes extraits est d'entre 3 et 20 phrases.

Appendix-2: Prompts used for Data Labelling

Les Prompts utilisés dans la phases de labéllisation: Les outils G-AI utilisés incluent Chat-GPT, Mistral, DeepSeek et Perplexity.

I. Prompt pour la labélisation des textes selon les intentions standards: "Dans les textes du fichier excel, nous voulons détecter des intentions liées au secteur de la santé. Il faut ajouter des colonnes binaires d'intention selon les intentions exprimées dans chaque texte: attribuer la valeur 1 si l'intention est exprimée dans le texte, attribuer la valeur 0 si l'intention est absente. Les seules valeurs acceptées sont les valeurs binaires 0 et 1 pour dire intention absente ou intention présente. Voici les règles:

1. Intention transactionnelle (Transactional): L'utilisateur est prêt à effectuer une action spécifique, comme acheter, s'inscrire, télécharger ou réserver. Cela reflète un comportement orienté vers un objectif immédiat. **Exemples:**

- "Prendre rendez-vous chez un cardiologue à Lyon"
- "Acheter un tensiomètre électronique en ligne"
- "Télécharger une application de suivi glycémique"
- "Réserver une téléconsultation avec un dermatologue"
- "S'inscrire à un programme de rééducation post-opératoire"
- "Commander des lunettes de vue sur internet"
- "Souscrire une mutuelle santé en ligne"
- "Acheter des vitamines pour enfants sur Doctipharma"
- "Faire une demande de devis pour une chirurgie dentaire"
- "S'abonner à un service de coaching nutritionnel"

2. Intention navigationnelle (Navigational): L'utilisateur veut accéder à un site, service ou page spécifique qu'il connaît déjà. Le moteur de recherche est utilisé comme un raccourci. **Exemples :**

- "Doctolib connexion patient"
- "Page d'accueil Ameli.fr"
- "Compte utilisateur Vidal Médicaments"
- "Portail de la Clinique Pasteur Toulouse"
- "Mon espace santé.gouv.fr"
- "Laboratoire d'analyses médicales Cerballiance"
- "Site officiel OMS coronavirus"
- "Page contact Hôpital Necker Paris"
- "Mon dossier médical CHU Lille"
- "Pharmacie Lafayette site officiel"

3. Intention informationnelle (Informational): L'utilisateur cherche à acquérir des connaissances ou à comprendre quelque chose, sans forcément vouloir agir ensuite. L'objectif est d'apprendre, d'explorer ou de s'informer. **Exemples:**

- "Quels sont les symptômes précoces du diabète ?"
- "Comment fonctionne une IRM ?"
- "Causes possibles des migraines chroniques"
- "Différence entre grippe et Covid-19"
- "Conseils pour mieux dormir naturellement"
- "Quels sont les effets secondaires de la chimiothérapie ?"
- "Comment prévenir l'hypertension artérielle ?"
- "Qu'est-ce que la télémédecine ?"
- "Quel est le rôle de la vitamine D dans l'organisme ?"
- "Pourquoi les enfants ont-ils souvent des otites ?"

4. Intention commerciale (Commercial): L'utilisateur effectue des recherches avant achat : il compare, consulte des avis, ou cherche des alternatives. Il ne souhaite pas encore acheter, mais il considère sérieusement ses options. **Exemples:**

- "Meilleurs glucomètres en 2025"
- "Comparaison entre pompes à insuline Medtronic et Omnipod"
- "Avis sur les prothèses auditives Siemens"
- "Fauteuils roulants électriques les mieux notés"
- "Comparatif des montres connectées de santé"
- "Quel purificateur d'air médical choisir pour l'asthme ?"
- "Orthèses de genou pour sportifs : avis et comparatif"
- "Cliniques dentaires abordables à Paris"
- "Quelle mutuelle santé couvre le mieux l'optique ?"
- "Appareils CPAP les plus fiables pour l'apnée du sommeil"

Règles à respecter:

- Une intention ne reçoit 1 que si elle est clairement exprimée et implicite, avec des actions concrètes ou mots/expressions associés.
- Ne pas supposer ou interpréter à partir d'indications faibles ou vagues. N'inférez pas au-delà de ce qui est littéralement dit ou fortement implicite.
- Si une intention n'est pas présente son score vaut 0.
- Si le signal est vague, absent ou simplement hypothétique, attribuez la valeur 0 (absente).
- Une phrase peut contenir plusieurs intentions ou au moins une intention.
- Chaque intention est évaluée indépendamment.

Les intentions sont intitulées en anglais *(Transactional, Navigational, Informational, Commercial)* dans le fichier excel joint dans les noms de colonnes vides à remplir par 1 (intention présente) ou 0 (intention absente). Il faut remplir les colonnes d'intention et fournir le fichier mis à jour.

II. Prompt pour la labélisation des textes selon les intentions de santé « Health Care Intents »: Dans les textes du fchiers excel, nous voulons détécter des intentions liés au secteur de la santé. Il faut remplir les colonnes vides selon les intentions exprimééss dans chaque texte : attribuer la valeur 1 si l'intention est exprimé dans le texte, attribuer la valeur 0 si l'intention est abscente. Les seules valeurs acceptées sont les valeurs binaires 0 et 1 pour dire intention absente ou intention présente. L'intention diagnostique est présente dans tous les textes. **Voici les règles:**

1. Intention de precaution: Rechercher des informations pour prévenir les problèmes de santé ou préserver son bien-être. **Exemples:**

- Comment renforcer naturellement son immunité
- Conseils pour éviter la grippe saisonnière
- Aliments pour booster les défenses immunitaires
- Exercices pour prévenir les douleurs lombaires
- Routines de sommeil pour éviter l'insomnie
- Habitudes saines pour préserver sa vue
- Comment éviter les infections urinaires
- Astuces pour réduire le stress au quotidien
- Précautions à prendre pendant une canicule
- Comment protéger sa peau du soleil

2. Intention diagnostique: Identifier la cause de symptômes ou de problèmes de santé. **Exemples:**

- Pourquoi ai-je une toux persistante ?
- Symptômes du diabète

- Causes possibles des maux de tête frequents
- Est-ce que mes vertiges sont inquiétants ?
- Comment reconnaître une allergie alimentaire ?
- Ballonnements fréquents : quelles explications ?
- Douleur à la poitrine sans effort : que signifie-t-elle?
- Picotements dans les mains : symptôme de quoi ?
- Fièvre sans autre symptôme : que rechercher ?
- Symptômes précoces de la maladie de Lyme

3. Intention thérapeutique: Trouver des solutions ou des remèdes à une affection diagnostiquée. **Exemples:**

- Meilleurs traitements pour l'arthrite
- Remèdes maison contre les migraines
- Que prendre contre une sciatique ?
- Traitement naturel contre l'acné
- Médicaments efficaces pour les reflux gastriques
- Huiles essentielles contre les maux de gorge
- Comment soulager une entorse à la cheville
- Que faire contre les aphtes douloureux
- Soins pour les peaux très sèches en hiver
- Physiothérapie pour hernie discale : est-ce utile ?

4. Intention de courtoisie: Rechercher des informations de santé pour autrui (soin, inquiétude, accompagnement). **Exemples:**

- Comment aider un proche souffrant de depression
- Meilleurs aliments pour une personne se remettant d'une operation
- Comment parler à quelqu'un qui refuse de se soigner
- Activités pour stimuler un senior atteint d'Alzheimer
- Conseils pour soutenir un adolescent en détresse psychologique
- Peut-on accompagner un proche à une séance de chimiothérapie ?
- Que faire quand son enfant refuse de manger ?
- Comment gérer l'anxiété d'un partenaire avant une operation
- Quel régime après une opération de la vésicule biliaire
- Vaccins recommandés pour les nourrissons

5. Intention d'urgence: Problèmes de santé immédiats nécessitant une action rapide. **Exemples:**

- Que faire en cas de crise cardiaque
- Soins d'urgence pour les brûlures
- Mon enfant s'est cogné la tête, dois-je m'inquiéter ?
- Comment réagir à une réaction allergique sévère
- Premier geste en cas d'étouffement
- Que faire si quelqu'un perd connaissance ?
- Quels sont les signes d'un AVC ?
- Urgence dentaire : que faire quand une dent se case ?
- Que faire après une morsure de chien ?
- Que faire si je me suis coupé profondément ?

6. Intention de se rassurer: Rechercher une confirmation ou une tranquillité d'esprit concernant sa santé. **Exemples:**

- Est-il normal de se sentir fatigué après une vaccination ?
- Effets secondaires des antibiotiques
- Fatigue après le COVID : combien de temps ça dure ?
- Est-ce dangereux d'avoir une tension basse ?
- Boule sous la peau : est-ce toujours grave ?
- Une fièvre légère peut-elle durer plusieurs jours ?
- Mes règles sont irrégulières : dois-je m'inquiéter ?
- Perte de cheveux après un accouchement : normal ?
- Taches blanches sur les ongles : signe de quoi ?
- Est-ce que les palpitations sont toujours un signe de problème cardiaque ?

Respecter ces règles:

- Pour un texte donné, nous voulons évaluer la présence de chaque type d'intention indépendamment en attribuant LE LABEL 1 POUR LA PRéSENCE ET 0 POUR L'ABSCENCE.
- Une intention ne reçoit 1 que si elle est clairement exprimée et implicite, avec des actions concrètes ou mots/expressions associés.
- Ne pas supposer ou interpréter à partir d'indications faibles ou vagues. N'inférez pas au-delà de ce qui est littéralement dit ou fortement implicite.
- Si une intention n'est pas présente son score vaut 0.
- Si le signal est vague, absent ou simplement hypothétique, attribuez la valeur 0 (absente).
- Une phrase peut contenir plusieurs intentions ou au moins une intention.
- Chaque intention est évaluée indépendamment.

Les intentions sont intitulées en anglais *(Precautionary, Reassurance, Diagnostic, Treatment, Urgency, Courtesy)* dans le fichier excel joint dans les noms de colonnes vides (à replir par 1 (intention présente) ou 0 (intention absente). Il faut replir les colonnes vides et fournir le fichier mis à jour.