1. Healthcare

Fraud Detection in Health Insurance

Problem Statement:
Health insurance fraud is a serious problem that affects both insurance companies and their clients. Fraudulent activities like over-billing, phantom billing, and billing for services that were not provided are common in the healthcare sector. These fraudulent activities increase the cost of healthcare for everyone and result in a decrease in the quality of care. Moreover, it puts the health and financial well-being of patients at risk.

Solution Type: Classification

Solution:
Using deep learning as a classification model against fraudulent health insurance providers can significantly improve the detection and prevention of fraudulent activities.
One way to use deep learning for fraud detection is to train a neural network to classify health insurance claims as either fraudulent or legitimate. The neural network can be trained on a large dataset of historical claims data, which includes both fraudulent and legitimate claims. The neural network will learn to identify patterns and features that are common to fraudulent claims and can be used to predict the likelihood of a new claim being fraudulent.
To improve the accuracy of the classification model, we can use a variety of deep learning techniques such as convolutional neural networks (CNN), long short-term memory (LSTM) networks, and autoencoders. These techniques can help to identify more complex patterns and dependencies in the data, and can also help to reduce the false-positive rate.

Dataset Link:
https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis?resource=download

Diagnosis of Medical Reports

Problem Statement:
Diagnosis of medical reports is a very important task in the medical field as it helps the doctors and patients realize the root of the problem and think of specific procedures to treat it. Now, fixing an appointment with a registered medical practitioner can be time consuming and sometimes detrimental to the survival of the patients in case of an emergency.Inaccurate or delayed diagnosis can have serious consequences for patient outcomes, leading to unnecessary treatments, prolonged hospital stays, and even death.

Solution Type: Text2Text Generation

Solution Statement:

To tackle the problem of accurate and efficient diagnosis of medical reports, we can leverage the power of deep learning and natural language processing (NLP) using a Text2Text generation model. Such a model can learn to understand and interpret medical reports, identify relevant clinical features, and generate accurate diagnoses and treatment recommendations.The model can be trained on a large dataset of medical reports, including clinical notes, and laboratory results, with each report labeled with the corresponding diagnosis. The Text2Text generation model can then learn to generate a diagnosis based on the input report, allowing it to accurately diagnose medical conditions and suggest appropriate treatments.To improve the accuracy and reliability of the model, it can be fine-tuned with transfer learning techniques, where the pre-trained model is fine-tuned with a smaller, more specific medical report dataset.
To ensure the model's safety and reliability, it can be validated and tested with an independent dataset of medical reports. The performance of the model can be measured using standard evaluation metrics, such as precision, recall, and F1-score, to assess its accuracy and performance.
So, using a Text2Text generation model trained with deep learning techniques and fine-tuning, we can improve the accuracy and efficiency of diagnosing medical reports, enabling better patient care and outcomes.

Dataset Link:
https://www.shaip.com/offerings/transcribed-medical-records-medical-data-catalog/

Flagging of potential infected people - Video Summarization

Problem Statement:

The identification and flagging of potentially infected people with COVID-19 is a crucial challenge for healthcare systems worldwide. As COVID-19 is highly contagious, identifying infected individuals is critical to preventing the spread of the virus. However, the current methods of flagging potentially infected individuals, such as symptom screening and contact tracing, have proven to be insufficient. Asymptomatic individuals, false negatives, and delays in testing and reporting further complicate the identification process. Therefore, there is an urgent need to develop and efficient methods of identifying and flagging potentially infected individuals, leveraging advances in technology and data analysis to improve pandemic management and control.

Solution Type: Video Summarization

Solution Statement:

To address the challenge of identifying and flagging potentially infected individuals with COVID-19, we can leverage the power of deep learning and video summarization models. These models can analyze large amounts of video footage from public places, such as airports, train stations, and other high-traffic areas, to identify individuals who exhibit symptoms or behaviors associated with COVID-19 infection.

The video footage can be analyzed using computer vision and machine learning techniques to detect and track individuals' movements and actions. This analysis can identify potential COVID-19 symptoms, such as coughing, fever, and shortness of breath, as well as other behaviors associated with COVID-19, such as touching the face or not wearing a mask.

To make this analysis more efficient, we can use video summarization models that can identify the most critical frames in the video footage and summarize them. The summarized video can then be reviewed by human experts who can further flag individuals who exhibit COVID-19 symptoms or behaviors, and take appropriate action, such as testing, quarantine, or contact tracing.

To ensure the accuracy and reliability of the model, it can be trained on a large and diverse dataset of videos, including different locations and populations. The performance of the model can also be regularly evaluated using standard metrics, such as precision and recall.

In conclusion, by leveraging deep learning and video summarization models, we can develop a more accurate and efficient method of identifying and flagging potentially infected individuals with COVID-19, helping to prevent the spread of the virus and better manage the pandemic.

Dataset:

1. Mall camera feeds data
2. Aadhaar Data set

Both are technically confidential datasets hence hard to acquire.

Detection of Kidney stones/tumor/cyst (s)

Problem Statement:
The accurate detection of kidney stones, tumors, or cysts is a crucial challenge for healthcare professionals due to the complex and diverse nature of these conditions. Kidney stones, tumors, and cysts can be challenging to detect, and symptoms may not always be apparent in the early stages. Delayed or inaccurate detection can lead to more severe health problems, such as kidney failure, infections, or cancer. Therefore, there is an urgent need to develop more accurate and efficient methods of detecting kidney stones, tumors, or cysts, leveraging advances in medical imaging and artificial intelligence to improve patient outcomes and reduce healthcare costs.

Solution Type: **Object Detection**

Solution Statement:
To address the challenge of accurate detection of kidney stones, tumors, or cysts, we can leverage the power of deep learning and computer vision using an object detection and image processing model. Such a model can learn to detect and identify features associated with these conditions, enabling earlier detection and better patient outcomes.

The model can be trained on a large dataset of medical images, such as CT scans, X-rays, or ultrasound images, with each image labeled with the corresponding diagnosis. The object detection model can then learn to identify regions of interest in the images that correspond to kidney stones, tumors, or cysts, enabling more accurate and efficient detection.

To improve the accuracy and reliability of the model, it can be fine-tuned with transfer learning techniques, where the pre-trained model is fine-tuned with a smaller, more specific medical image dataset. This approach can help the model adapt to specific medical specialties or sub-specialties, such as urology, radiology, or oncology.

To ensure the model's safety and reliability, it can be validated and tested with an independent dataset of medical images. The performance of the model can be measured using standard evaluation metrics, such as sensitivity, specificity, and accuracy, to assess its performance and reliability.

In conclusion, by using an object detection and image processing model trained with deep learning techniques and fine-tuning, we can improve the accuracy and efficiency of detecting kidney stones, tumors, or cysts, enabling earlier diagnosis and better patient outcomes.

Dataset Link:
https://www.kaggle.com/datasets/nazmul0087/ct-kidney-dataset-normal-cyst-tumor-and-stone

2. Politics

Election Result Prediction

Problem Statement:
The widespread use of the social media network for communication, social participation, and self-expression has resulted in many individual data being willingly disclosed online. The social disclosure on social media provides a platform for the researchers to measure general opinion unremarkably. Computer technology-based social media platforms have benefited people for self-expression, communication, and social participation. These social media sources are responsible for sharing text, audio, and video links between individuals worldwide to be well informed and keeping in touch with others. This explosive social development can be seen on Facebook, Linked In, FriendFeed, Myspace, and many more, but the flow of Twitter is much more than other platforms. Many people have been attracted by social media platforms such as Instagram, Twitter, and Facebook, which allow them to express their thoughts, feelings, and ideas about various people, places, and things. People give their views on these social websites for the political party election campaign.

Solution Type: **Classification**

Solution Statement:
The proposed election results prediction framework predicts Twitter opinions of the general election. The tweets were posted about current trends of the political party, which users in hashtags consider to express their views. The received tweets are stored in the database, and the dataset is pre-processed. After pre-processing, the resting dataset is divided into the testing and training set. Deep learning is then applied to the training set in RapidMiner, and a classification model is established. The model is tested on the testing dataset, and its accuracy is compared with state-of-the-art schemes.

The intention of using multiple machine learning algorithms like Naive Bayes, SVM, and deep learning is to compare their accuracy. RapidMiner is used as a simulation tool that has predefined operators and classifiers.

We focused on determining the approach to adopting different machine learning algorithms with the maximum accuracy learning rate about election sentiment. In documents, words of many semantic orientation sentences or phrases are calculated using lexicon-based sentiment analysis. However, to classify semantic orientation sentences for phrases or text, algorithms such as SVM, Naïve Bayes, and deep learning techniques are used. We investigated the predictive power of Twitter using Sentiment analysis.

Dataset Link:

https://assets.researchsquare.com/files/rs-839553/v1/8285e1d5-4690-43f5-b123-a20fb61e037f.pdf?c=1638373775

Speech analysis of politicians

Problem Statement:

The analysis of politicians' speech is an essential aspect of political discourse, providing insight into policy proposals, messaging, and overall political strategy. However, analyzing and summarizing a politician's speech can be a daunting task, especially for long speeches or debates. This can lead to a lack of clarity and transparency in political communication, making it challenging for voters to make informed decisions. Therefore, there is a need for an automated solution that can analyze and summarize politicians' speeches, highlighting the essential points of their speeches, and making it easier for voters to understand and evaluate their policy proposals.

Solution Type: Summarization

Solution Statement:

To solve the problem of summarizing politicians' speeches, we can utilize deep learning summarization models. These models can analyze a politician's speech and generate a summarized list of essential points, allowing voters to easily understand their proposals.

The deep learning model can be trained on a large dataset of speeches, debates, and other political events, using unsupervised or supervised learning techniques. The model can then use natural language processing (NLP) techniques to analyze the text, identify key phrases and summarize the main ideas presented.
To ensure the accuracy and quality of the model's summaries, it can be fine-tuned with transfer learning techniques. This can involve fine-tuning the pre-trained model with a smaller, more specific political speech dataset, ensuring the model can adapt to different political contexts and speech styles.

To ensure that the model generates high-quality summaries, it can be evaluated and validated against independent datasets of political speeches, with metrics such as Rouge-1, Rouge-2, and Rouge-L used to measure the quality of the summaries.

In conclusion, by utilizing deep learning summarization models, we can automatically analyze and summarize politicians' speeches, making it easier for voters to understand their proposals. With fine-tuning and validation techniques, we can ensure that the model generates high-quality summaries, leading to a more informed electorate and a more transparent political process.

Dataset Link:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/R59EUT

Fake News Detection:

Problem Statement:

The spread of fake news and propaganda has become a significant problem in modern times, with the potential to cause harm and damage to individuals and society. Analyzing the text and images used in these fake news stories can help identify patterns and common features that distinguish them from legitimate news. However, detecting and flagging fake news in real-time can be a challenging task, especially with the increasing volume of information shared through social media and other channels. Therefore, there is a need for an automated solution that can accurately identify and flag fake news stories, enabling users to make informed decisions and preventing the spread of misinformation.

Solution Type: Text2Text Generation

Solution Statement:

To address the problem of fake news detection, we can use a deep learning text2text generation model. The model can be trained on a large dataset of news articles, both legitimate and fake, allowing it to learn to identify patterns and common features that distinguish the two types of articles.

The text2text generation model can generate summaries of news articles, highlighting the key points and identifying any inconsistencies or biases in the article. It can also compare multiple news articles on the same topic, identifying any differences or similarities between them.

To further enhance the model's accuracy, it can be fine-tuned with transfer learning techniques. This involves taking a pre-trained model and fine-tuning it with a smaller, more specific dataset of news articles, allowing it to adapt to specific domains and topics.

In practice, the text2text generation model can be integrated into a web browser extension or a mobile application. Users can simply input the URL or upload the text of an article they are unsure about, and the model will analyze the article and generate a summary, highlighting any potential issues or inconsistencies. This can help users make informed decisions and prevent the spread of fake news and propaganda.

Overall, a text2text generation model provides a powerful tool for detecting and combating the spread of fake news, ensuring a more informed and responsible use of news sources.

Dataset Link:
https://www.kaggle.com/c/fake-news/data?select=test.csv

Predicting Policy Domains from Party Manifestos:

Problem Statement:

During campaigns, political actors communicate their position on a range of key issues to signal campaign promises and gain favor with constituents. Identifying the political positions of political actors is essential to understanding their intended political actions. This is why policy preferences—or positions on specific policy issues expressed in speech or text—have been extensively analyzed within the relevant political science literature. Methods employed to investigate the policy preferences of political actors include analysis of roll call voting, position extraction from elite studies or regular surveys, expert surveys, hand-coded analysis, and computerized text analysis. Studies that utilize political manifestos, electoral speeches, and debate motions often rely on the availability of machine readable documents that are labeled by policy domain or policy preference.

Solution Type: Speech to Text

Solution Statement:

Hand-labeled political texts are often required in empirical studies on party systems, coalition building, agenda setting, and many other areas of political science research. While hand-labeling remains the standard procedure for analyzing political texts, it can be slow, expensive, and subject to human error. Recent studies in the field have leveraged supervised machine learning techniques to automate the labeling process of political texts. We build on current approaches to label shorter texts and phrases in party manifestos using a pre-existing coding scheme developed by political scientists for classifying texts by policy domain and preference. Using labels and data compiled by the Manifesto Project, we make use of the state-of-the-art Bidirectional Encoder Representations from Transformers (BERT) with Convolutional Neural Networks (CNN) and Gated Recurrent Units (GRU) to seek the best model architecture to supplant manual coding of political texts. We find that our proposed BERT-CNN model outperforms other approaches for the task of classifying political texts by policy domain. There are several avenues for future work on neural networks and deep language representation models for the automatic labeling of political texts. For instance, investigating the features of individual categories that demonstrate superior performance could shed light on how we could incorporate additional features of texts to improve model performance. This area of research would also benefit from better understanding how we can filter out texts that do not fall into a particular classification scheme. Knowledge on how these issues could be resolved to improve model performance would allow for extensions in the application of deep learning models to the classification of political texts.

Dataset Link:

https://stagezero.ai/data-sourcing/