CHAPTER 7

Potential Use Cases of Algorithmic Government

Abstract This chapter presents various potential use cases that can be developed in different domains, specifically for developing and underdeveloped nations. Multiple codes for prototypes of use cases from this chapter are presented for technology enthusiasts. The problem of fake news in media, managing floods in disaster management, assessment in the education system, the online admission process for schools and colleges, loan fraud issues in finance industry, and online reputation management are some of the use cases discussed in the chapter which can be automated for the Government sector. Although these are the early stages of algorithmic implementation in various regions, with the amount of data increasing, governance can be made more robust and flexible.

Keywords Education · Media · Disaster management · Reputation management · Finance industry

7.1 Media

The Problem of Fake News

Fake news depicts sensationalism, which presents misinformation or hoaxes that spread through traditional media and popular online social
media. The rapid increase in social media has changed the way of acquiring information. More news is consumed through social media, providing information timely and comprehensively on the events worldwide. As compared to traditional communication (newspapers or television), visualized info like images and videos explain better and attract attention from the viewers or the readers. With misleading words, social network users get easily affected by fake news, which leads to tremendous effects on offline society. Reading a story on social media is becoming easier and accessible, due to which fake news becomes a significant issue for the public and the government. Fake news misleads readers and spreads, which negatively affects or manipulates public events. Fake news creates a lot of hurdles for the government at every level. During this pandemic, authorities are found busy in clarifying people to protect them from rumors.

The reasons for people switching from traditional to online news consumption, i.e., advantages of online use of news are that it is less expensive as compared to conventional media; requires less time than on traditional media; and it is easy to share content in the form of videos, blogs, or pots with other users or comment on social media. There is also a disadvantage of social media, i.e., the quality of news on social media is lower than the traditional news. As it is cheaper, faster, and easier to provide news online, much fake news is produced for many reasons, such as financial and political gain. This gives ease to publishers to publish their articles in collaborative environments. People believe that the information received from social media sites is reliable. It is analyzed that people are unable to recognize deception, which affects the news ecosystem.

There are various psychological factors due to which people believe in fake news: (i) social credibility, i.e., people think the source is credible if others find the source if credible, especially when there is not enough information about the source; (ii) frequency heuristic, i.e., people favor information that they frequently hear even if it is fake. Many malicious accounts on social media become sources of fake news. There are mainly five types of fake news—first, Deliberate Misinformation, which is information that is spread in a way to deceive target users. Second, Clickbait grabs the user’s attention so that the user clicks on the fake news. Third, Parody or Satirical articles use absurdity and exaggeration to comment on events that unease the readers. Fourth, False headlines use to draw the attention of the reader. The title may not match the context of the
article. This type is untrue and misleading. The fifth is Hoaxes, which deceives the reader by causing harm and material loss to the user.

Why Fake News Spread Is a Concern?

The spread of fake news has a severe negative impact on society. Firstly, fake news destroys the authenticity balance of the news ecosystem. For example, it can be seen that fake news is more widely spread than popular authentic news. Secondly, fake news makes people accept false beliefs. Lastly, fake news changes how people interpret and respond to the real story. For example, some fake news is created just to trigger people’s trust and confuse them, hinder their ability to differentiate between what is right and what is not valid. Fake news many times leads to communal riots. For example, some lawbreakers spread a video a few years ago on WhatsApp and other social media platforms, which lead to the enormous communal violence that led to the destruction of belongings to different communities.

The spread of fake news containing forged or misinterpreted images can cause many adverse effects like manipulating important events. For example, recently, there was a piece of fake news that depicted an entire community of people as a disease source (COVID-19). A few people in mid-March at a religious place in Delhi led to many positive cases in India. Many fake videos depicted the group as ‘Corona Villains.’ A video was spread that claimed some foreign country people intentionally licked kitchen utensils to spread the novel coronavirus. The footage was fact-checked, which proved that a group of humble people consumed food from those utensils to ensure that no grain of food was left and wasted. During this pandemic, authorities are found busy in clarifying people to protect them from rumors.

According to a statistical report, 47% of the population acquires news from online mediums like social media, and others receive from TV, newspapers, and radio. The most trusted news source in 2018 was newspapers. The primary source of news for people is an online search engine, and 45% of the population believes the story is accessed via search engines. Top networks used to access reports are Facebook and WhatsApp by 52% of the community. 41.8% of fake news data in the election is due to social media, which is higher than shares of both traditional mediums (TV/radio/print) and online search engines.
Therefore, it is essential to help reduce the adverse effects caused due to fake news such that it benefits both the public and the news ecosystem.

**How Algorithms Can Provide Solution**

The steps performed in the implementation of Fake News Detection are mainly data preprocessing, data visualization, feature extraction, and applying models. The following are the steps.

1. Create a column ‘label’ in the data frame, which is our target feature denoting whether the news is True or Fake.
2. Combine the columns ‘title,’ ‘text,’ and ‘subject’ into one feature column called ‘article.’
3. Create the final data frame with only ‘article’ and ‘label’ features.
4. Start with data preprocessing, i.e., transform the raw data into a useful and efficient format like removing null values, punctuation, stop words, and then perform lemmatization.
5. Search for the null values and removing them from the data frame if present.
6. Convert every word in the ‘article’ feature to lower case.
7. Remove the punctuation from the data, i.e., full stop, colon, comma, brackets, etc.
8. Remove the stop words, i.e., useless words in the data like ‘a,’ ‘an,’ ‘the,’ ‘in,’ etc.
9. Perform lemmatization, i.e., the process of grouping together different forms of a word to its root form such that they can be analyzed as a single item.
10. Then perform the data visualization, i.e., bar plots, pie charts, word cloud, etc.
11. The next important step is feature extraction, i.e., attribute reduction in the dataset.
12. Create a bag of words (BOW), i.e., text (sentence or document) is represented as a bag (multiset) of words, ignoring grammar but keeps count of word occurrences.
13. Create TF-IDF (Term Frequency–Inverse Document Frequency) vectorizer, i.e., mathematical statistic that reflects how important a word is to a document and contains information on both more essential and less important words. It converts the collection of data into a matrix of TF-IDF features.
14. Then comes the final phase, which comprises of applying various models like Passive-Aggressive Classifier, Naïve Bayes, Stochastic Gradient Descent (SGD), Logistic Regression, Random Forest Classifier, K-nearest neighbors classifier, etc.

15. Applying the classification model has initialized the model, trains the classifier, predicts the target, and then finally evaluates the classifier model.

The block diagram of the process is shown in Fig. 7.1.

Which Analytical Techniques Can Help?

The text classification problem requires to define three sets. First, is training dataset \( D = \{d_1, d_2, \ldots, d_n\} \), second is the class label \( C = \{c_1, c_2, \ldots, c_n\} \) and last, is the test dataset \( T = \{d_1, d_2, \ldots, d_n\} \). Each data
"di" of the training dataset is labeled with the class label "ci" from the class label set C, and each data of the test dataset is unlabeled. The classifier’s main aim is to construct a model from training data by relating features to one of the target class labels. After the classification model is trained, it can predict the class labels of test data. The formula for both training and testing is given in Eq. 7.1 and 7.2.

\[ f : D \rightarrow C \]  
(7.1)

\[ f(d) = c \]  
(7.2)

Preprocessing generally involves steps such as stop-word removal, punctuation removal, lemmatization, etc. Feature extraction is a data representation process that includes several activities to scale down data complexity and carry out the classification process in an accessible manner. It involves the calculation of TF (Term Frequency) and IDF (Inverse Document Frequency) from tokenized data. Finally, the data is normalized to unit length to perform classification efficiently.

F. Sebastiani surveyed the different classifications and discussed the specific role of machine learning algorithms in classification. Colas et al. compared SVM’s performance with KNN and Naïve Bayes classification algorithms on Reuters 21,570 dataset. The performance metrics were calculated for each classifier. The overall performance of the SVM classifier worked well. If the preprocessing and internal parameters are adjusted well, then KNN and Naïve Bayes perform well too.

In logistic regression, the probabilities of the outcomes are modeled using a logistic function. This model is designed mainly for classification and helps understand the effect of independent variables on the dependent or outcome variable. It works only when the predicted variable is binary and assumes there are no missing values.

Naïve Bayes classifier is a probabilistic classifier that works well in real-world situations like document classification and spam filtering. It requires a small amount of training data to predict. This classifier is extremely fast as compared to others. There are two variants: the Multivariate Bernoulli model (B_NB) and Multinomial model (M_NB). Bernoulli works well on binary data. The multinomial model works on frequencies of attributes available in vector space representation of data.
Stochastic Gradient Descent is an efficient and straightforward classifier to fit linear models. It is helpful when the number of samples (data) is large enough. It is sensitive to feature scaling.

K-nearest neighbors (KNN) is a lazy learning model, stores instances of training data. Classification is done from a majority vote of \( k \)-nearest neighbors at each point. It is simple to implement, robust to noisy training data, and effective when the number of samples or data is large but has high computation costs.

Random forest classifier fits the number of decision trees on subsamples of datasets (size of subsamples is the same as original input sample size; these samples are drawn with replacement) and uses an average to improve the prediction accuracy. It is more accurate than a decision tree and removes overfitting, but it is complex and challenging to implement the algorithm.

Passive-Aggressive algorithm remains passive when the classification outcome is correct but becomes aggressive when miscalculation (updating and adjusting).

These classifiers’ efficiency is measured by performance metrics like accuracy, precision, recall, F1 score, etc. The highest accuracy for the dataset used in this book (Appendix A) is of Passive-Aggressive classifier, which is 99.5%, and next is that of Stochastic Gradient Descent, which is 99.1%. Passive-Aggressive classifier has the highest F1 score. Therefore, the classifier which can best fit the data is the Passive-Aggressive classifier. The prototype code with algorithmic implementation is shown in Appendix A.

**Conclusion**

The extensive spread of fake news has plenty of negative impacts on society. The false news detection system can help mitigate the adverse effects of fake news on both the community and the news ecosystem. The solution can maintain the authenticity balance of the news ecosystem. It can help people accept false beliefs and differentiate between what is right and what is not. Fake news is not a problem; the rise of this fake news on social media makes it sturdy and challenging. The rate of production of counterfeit digital news is large and rapid. Thus it is challenging for machine learning to detect fake news effectively. False news detection can help to reduce misinformation risk and communal riots.
Using fake news detection, alert systems can be created that can alert users whether the news or the page is fake or not. Also, the alerts can be sent to the media so that they can inform the public. Various detection systems can be created as an extension of this. Also, intervention systems can be designed that can intervene to offload the page if it is fake. Malicious accounts containing false news can be removed so that the spread is reduced. The primary source of the news can be determined, and that source can be further blocked to stop the spread. Users with certain news can be immunized rather than fake ones. It can help minimize the spread scope and reduce fake news that created unnecessary tension in the polity. It can reduce public chaos and censorship over the media. It is a tool that can help editors and journalists to pace up the verification procedure for the content that has been generated from social media. And the quality of the news can be improved.

7.2 DISASTER MANAGEMENT

The Problem of Floods as Natural Disaster

A country like India is a sub-continent located in the core of the summer monsoon belt. India has many major river systems, such as the Himalayan Indus Ganga Brahmaputra system and the peninsular Godavari, Mahanadi, Krishna, Kaveri on the east coast, and Narmada and Tapi on the west coast. After Bangladesh, India is most prone to floods. The most vulnerable states are the coastal regions which are Gujarat, Maharashtra, Goa, Karnataka, Kerala, Tamil Nadu, Andhra Pradesh, Orissa, and West Bengal. It receives more than 80% of rainfall from June to September, during which most floods occur. India faces floods almost every year with varying magnitudes. Tide is the most persistent disaster. Floods are mainly caused by riverbanks’ low capacity to control high flow bought down from the upper catchment caused due to heavy rainfall. In coastal areas, floods are caused by cyclones or typhoons. Flash floods occur in low lying areas near foothills. Some other causes of floods are backing up water in tributaries at their outfalls in the main river with landslides blocking the stream, resulting in backwater overflowing riverbanks. The leading cause of the flood is heavy or excessive rainfall that occurs mainly in monsoon months from July to September. Sometimes, floods are caused by Glacial Lake Outburst Floods (GLOFs), i.e.,
the surge which occurs when a dam consisting of a glacial lake fails to work.

Heavy rainfall, low capacity of rivers to carry extreme flood discharge, inadequate drainage to take away rainwater to streams or rivers, man-made factors like failure of dams are the major causes of flood. Also, ice jams or landslides blocking streams, backwater, debris flow, and cyclones cause the flood. Floods occur almost every year in some parts of the world or the other. Different regions have different climates or rainfall patterns in a country. Some regions suffer devastating floods, and some suffer drought. Due to varying rainfall, areas which are never prone to floods traditionally also experience it. Geomorphological, extreme rain, inflow, and outflow limits of sewerage also contribute to cascades. Indian Summer Monsoon (ISM) has different phases, which are—onset (mid-May to mid-July), peak rainfall (July to August), and withdrawal (mid-September to mid-October). Also, the shift of heat source over Northern India due to shifting of Intertropical Convergence Zone (ITCZ) generally stays between 20° and 30° North during peak rainfall months (July to August). Their shifting affects the monsoon due to which heavy rainfall occurs during this phase. Heavy rains are linked with heavy downpours, low pressure, and monsoon breaks. In the Himalayas, low pressure, Western Disturbances (WDs), and ISM lead to flood. For example, on 16–17 June 2013 Uttarakhand flood, GLOF from Mahatma Gandhi Sagar due to glacier/melt and glacial moraines led to an increase in overflow toward Kedarnath and downstream.

Deforestation also plays a vital role in the flooding equation. As trees prevent runoff, lower the number of trees, more is the water flow, leading to destruction. Therefore, the more is the forest cover, the lesser is the risk of flood.

**Why Flooding Is a Problem?**

Floods cause loss to both life and the economy to a large extent. According to Rashtriya Barh Ayog (National Commission on Floods), India has a geographical area of 329 Million hectares, out of which 40 million hectares (one-eighth of total area) are prone to flood. Every year, 7.5 Million hectares of land are affected, 1600 lives are lost, and other damages like crops, houses, and public utilities are of USD 250 Million. The highest loss of lives was 11,316 in 1977. Flood causes fear and insecurity in the minds of people residing in floodplains. The aftereffects of
floods are suffering from people, the spread of some diseases, unavailability of essential commodities and medicines, etc. High rainfall causes a rise in water level, which leads to submergence of regions, landslides, waterborne diseases, etc. Floods during the Rabi and Kharif seasons affect the food security of India.

For example, on 16–17 June 2013, the Uttarakhand flood led to massive destruction of life and land. Heavy rainfall caused flashflood in the Kedarnath valley, Mandakini, and Saraswati rivers. Interaction between West-Northwest moving monsoon, low pressure, and eastward-moving mid-latitude WD leads to extreme rainfall. During the same time, a monsoon low developed over the north Bay of Bengal and moved west-northwest across northern India. Low pressure formed on the Gangetic plain caused high moisture over the northwestern part of India.

Due to floods, salt in seawater contaminates land and reduces crop yield, electricity supply cuts, increased traffic congestion, increased costs of emergency services, loss of exports, etc.

To manage floods, the water level should be controlled, done by dams/reservoirs storing floodwater, to be leveled without exceeding in downstream. The lower the level of the dam, the lower is the risk of flood. Another method is levees, which consist of earthen dams built between rivers and the areas to be protected. When these methods fail, sandbags or portable inflation tubes are used. Another technique is dike; it lowers the risk of flood and protects land naturally underwater most of the time. The weir is also constructed across the river to change its flow, which acts as an obstruction for water flow.

**How Can Algorithmic Solution Help?**

A dataset with monthly rainfall amounts for each year from 1900 to 2015 for each state has the potential to explore the rainfall patterns, with which a prototype has been developed shown in Appendix B. The basic approach for the problem is that of a binary classification. Dataset also has the data of the duration of 3 months. Using the dataset, first data preprocessing is performed, like removing null values, which are zero in the dataset. Next, the average rainfall for every ten days is calculated. This intermediate data is given as an input to the machine learning model, which provides output labels as 0 or 1 (whether flood will occur or not). Next, the model is trained depending on a threshold value of average rainfall in the dataset. Giving the average data of 10 days to the model as input, the
model predicts whether there are chances of flooding or not by setting a threshold value in the training data.

There is an official website of government called cwc.gov.in, which updates itself regularly, telling whether an area’s water level is (i) normal flood, i.e., the situation at any flood forecasting sites when the water level of the river is below the warning level; (ii) above normal flood, i.e., if the river’s water level at the flood forecasting site touches or crosses the warning level but remains under the site’s danger level. This category is assigned a yellow color; (iii) severe flood, i.e., if the river’s water level at the forecasting site touches or crosses danger level but below the highest flood level of forecasting. Orange color is assigned to this category; and (iv) extreme flood, i.e., when the water level of river touches or crosses the highest flood level recorded by any forecasting site so far. This category is assigned a red color.

**Which Analytical Techniques to Use?**

The technique used for flood forecasting is logistic regression. This is because the overall aim was to get 0 or 1 labels as output, i.e., binary outcome, 0 for no severe risk of flood, and 1 for severe risk of flood. Linear regression can also be used for binary outcomes. But if it is used, it is possible to have fitted regression where predicted values for some individuals are outside 0 or 1 range of probability. For a binary outcome, logistic regression is used to model the log odds as a linear combination of parameters and regressors. Logistic regression is the most popular technique for classification and regression. It is used when probabilities between the two classes are required. In linear regression, the outcome can be any continuous value, making it difficult to classify, whereas logistic gives only binary value that is biased and with low variance. Various research papers show how logistic regression is better than linear regression for binary outcomes. V. V. Srinivas shows how fuzzy means clustering can be worked upon for flood prediction. A. Mosavi describes different machine learning models for flood prediction. Trace Smith also reveals that logistic regression is better than random forest classification for binary outcomes.
Can There Be an Impact of Technology?

The approach can be used to identify whether an area may flood or not, given the region’s rainfall data. The solution can be augmented with the existing solutions of dividing an area into different zones based on the severity of the floods or being used for alert predictors. If accessed, the database can be used to send a red alert warning to mobile phones of people living in the affected area using a free online text message sending portal. The people in the affected area can migrate to some other area so as for their protection. The government can take steps to protect the animal species too. Early warning systems can help a lot. Lives can be protected, and essential commodities can be collected at times by those who cannot migrate. Automated demand predictions and supply chain management can be done on the basis of flood predictions in advance.

7.3 Education & Assessment

The Problem of Transforming to Online Education

Online learning may have been on the rise in recent years. However, it has become a necessity and a temporary substitute in many places, which are still under extended lockdowns amid the COVID-19 pandemic. As education has undergone a tremendous transformation with the emerging technologies enabling digital learning, a key aspect may be to explore the potential of a segment that has not found large-scale adoption yet, i.e., the virtual exams. Exams are a crucial component of any educational program or a system, and online education programs are no exception. Given any exam, there is a possibility of cheating, and therefore its detection and prevention play a critical role. With several exams being postponed or conducted amid immense health risks to students, remote proctoring systems can enable them to take exams in the safety of their homes. What adds to this is, with AI-based invigilation technologies ensuring that they do not cheat or indulge in unfair means during the assessment, educational institutions can also benefit from this arrangement. The most critical factor among faculty and student concerns are student privacy and increasing test anxiety via a sense of being surveilled. Some experts might also argue that the whole premise of asking students to recall information under pressure without access to their course materials is flawed.

Online human monitoring is a common approach for proctoring online exams. Still, the main downside is that it is very costly in requiring
many employees to monitor the test takers. With AI being able to tackle the offline model’s challenges, especially amid such crucial times, it can prove to be a gamechanger and prepare students and educators for the post-pandemic world. Online exam conduction with AI-based proctoring can significantly reduce the human efforts and efficiently conduct the examinations online and proper monitoring of any suspicious actions or movements of the students. The main goal is to develop a proctoring system and eliminate the need for human proctors, reducing the cost, time, and the chances of any human errors in detecting any suspicious activities. The AI-based proctoring system is a real-time system that can completely manage the conduction of exams, from verifying the student’s identity to notice any unnecessary movements and generating warnings till the end of the exam.

The conduction of exams through an online mode and using AI-based systems used for proctoring would be an emerging trend, which can be quickly adopted in the times of a pandemic like COVID-19. This advancement in technology can prove to be a boon to any country throughout the world and to any institution or body, which cannot follow the traditional trend of taking the exams physically. Cutting the labor costs would prove to be technological advancement with higher reliability and accuracy than the conventional method of taking exams. While the challenge could be of the prerequisites for implementing this new advancement, of the laptop and a good internet connection, but once met, it can be of immense importance to certify or test any student taking an exam for his/her skills.

**Importance of Online Assessments**

Online learning is growing at a rapid rate. Due to lockdown and health crisis in the current situation, it is the need for the hour to build a system that makes our education system up and running. Online Exam conducting using AI proctoring system has become so comfortable using this. It is such an efficient and cost-effective way to conduct exams. Since assessments are a crucial component of any education system, it is necessary to conduct exams, and AI proctoring system can help to continue the exam process. The main concern when conducting online exams is whether the exams will be done without cheating and monitoring students. So, to solve this problem, the AI proctoring system keeps the camera switched on, and using artificial intelligence, each student can
be controlled individually. Moreover, to add functionality to this system, the system can give warnings, and after a specific number of warnings, the artificial proctor will cancel out the exam.

This system provides easy management for the institutions and reduces human error and effort. The labor cost is reduced using this remote AI proctor. AI proctor will itself see any suspicious activity and raise a warning on such a situation. AI proctor will minimize human error, and continuous monitoring of each student is not possible with physical invigilators’ help, but AI proctor manages and handles them with ease. AI proctoring is a scalable solution and can be implemented at various stages and a variety of platforms. AI proctoring system has been a technological boom to the education system, with such advancements in technology. This is very useful for the education system’s future, which is to set up its foot in the online education system. This solution helps those scenarios where the individual cannot reach the destination of the exam center. Using AI proctoring technology, students can give their exams online sitting at their home. Institutions need not worry about cheating or suspicious stuff as the system itself will manage that. This system is a convenient way to conduct exams in such a health crisis where leaving homes will jeopardize thousands of lives. Seeing the current situation and analyzing the fact that conducting exams is a necessity AI proctoring system seems to be the most feasible and reliable solution.

Understanding the health and education AI proctoring system’s concern bridges the gap between the two without compromising with anyone of them. This solution could result in a drastic and better change in the future. This system can set up an example for future how online exams can be conducted and changes that are required for the betterment of this system. Machine learning and Artificial Intelligence provide a platform to build such a scalable solution for this complex problem. AI proctoring system will reduce the cheating and analyze the marks accordingly based on the evaluation. The candidates capable of qualifying will only qualify using the AI proctor.

**Algorithmic Solution to Proctoring**

The AI proctoring system is built to solve remotely conducting exams, such as health crises, and managing those exams free from any cheating and continuous monitoring.
The examination problem to conduct at remote places online was a challenge. This system performs the exam remotely and online and ensures there are the least chances of the candidate using fraud methods to pass an exam.

Figure 7.2 shows the AI proctoring system overview, which starts using object detection, counting the number of persons, and tracking eyes and mouth to avoid cheating. If the mouth is opened, alerts are displayed on the screen, and analysis is done using the behavior of fraud of the candidate. Detect the eyes, mouth, and face of the person giving the exam. AI proctoring system tracks the movement of eyes, mouth, and face simultaneously. Using the analysis of the face’s parts, one can raise a warning for suspicious activity.

Figure 7.3 displays the face detection algorithm and overview of the prototype solution present in Appendix C. YOLO face detection first detects the persons given a frame of the image and find whether there is any person in the image or not. The probability of the face is calculated, and using anchor boxes, the person’s face will be outlined using a rectangular table. It then calculates if any other person is there in the frame. Using a sliding window algorithm, it then alerts the screen more than one person detected and displays an alert.

The algorithm, after detecting the face, now captures the points on the eyes and mouth, as per the given process in Fig. 7.4. Using those points on the eyes, it analyzes the movement of the eyes’ pupil. It displays the eye’s position wherever it moves, i.e., either left, right, or upwards if it detects any suspicious movement of the eye.

**Fig. 7.2** Flow diagram of AI proctoring solution (Source: Author creation)
Next up, it captures the position of the mouth. The data points obtained during face detection, the aspects of the mouth are analyzed. If the mouth is open to a certain threshold, it displays that the person has opened mouth, indicating the person talking to someone, saying this message on the screen, and sending an alert.

Analysis and display of alerts: The algorithm detects the person’s face and tracks the eyes and mouth. After all the detection and recognition, the alerts must be displayed. So, any suspicious activity including more than one person on screen or detection of remote or mobile phone on the screen, movement of eyes where the eyes move will be captured, and alerts will be displayed on the screen in real-time. If the person opens his/her mouth, the message is displayed as a proctor finds this as a way of cheating. The statement above is displayed on the screen.

The solution presented for conducting exams is a model that is ready to use in real-time environments while facing all the challenges that are offered toward it. It is robust to handle such situations.

**Analytical Techniques for AI Proctoring**

Object Detection: The principal technique used in the AI proctoring system is object detection and recognition. Object detection plays an important role when it comes to analyzing real-time data. An Online AI proctor needs to detect a person’s identity; it needs object detection. It should ensure that human is only giving the test and not any robot. Secondly, it must provide the status of the person passing an exam. Object detection includes this. It must also detect any object that is brought
during the exam, which is not allowed. For example, a mobile phone, earphones, and other such devices are not permitted during the exams. Object detection, for the solution in this book, seemed to work well. The reason behind selecting such a technique was to ensure the person’s real identity and use of any suspicious object that is not allowed during examinations. Object detection algorithm can detect many other things in the image in real-time that can be counted as suspicious.

Face Detection: This technique plays a significant role in detecting faces and the number of persons in the screen frame. As AI proctoring system uses the camera of the course being used, it must recognize that remotely some other person is not helping the candidate. So, our model needed to detect the number of persons. For the face detection algorithm, we used the YOLO algorithm. Using this, we counted the number of faces shown on the image, and a warning is indicated if more than a single person is detected in the camera. The face detection algorithm technique was our choice to discover the number of faces given a frame quickly. In face detection, we noticed two parts of the face—eyes and mouth. The discovery of the pupil of the eyes will tell whether the individual is looking on the screen or somewhere else. The mouth part detection will help in finding whether he opened his mouth for saying something or not. If the algorithm detects whether the person is opening his/her mouth, the algorithm will display a message on the screen. Using the detection of eyes and mouth, we can make sure that the person is not involved in any suspicious activity or cheating.

Object Tracking: this technique helps in measuring and tracking the objects we detected earlier. The eye-tracking or pupil tracking will track the area where the pupil of eyes moved, and accordingly, the proctor will display a message on the screen whether the individual is looking toward the left-right, bottom, or up. Next up for tracking, we use mouth tracking and lips tracking. If the lips are apart from each other, the tracking algorithm tells if the person is speaking or not. That can be used to verify if the person is involved in cheating. Object tracking algorithm helps detect and analyze the candidate’s cheating behavior, and that can be used in the marking scheme. We chose object tracking as this is the best way to solve our problem and conduct exams remotely.
Impact of AI Proctoring on Online Education

With this lockdown period growth of online education is quite a trend. So, conducting online exams is quite a challenging task. The solution above is going to have a considerable impact on the education system. The online conduction of exams based on the AI proctoring system is both beneficial and efficient. Building such a model is advantageous and robust as well. AI proctoring system can replace physical teachers, and cheating can be reduced to minimal. The quality of education will tend to improve as those students who are capable will qualify.

This system can detect the number of persons given a frame. If a person uses a cell phone, it will see that, and if a person is looking somewhere else apart from the screen, it will display the message and alert accordingly.

There is a significant variation in proctored and physical conduction of exams. It is challenging for a single teacher to look at all students, while this problem is easily solved using the proctor. The AI proctor will automatically recognize the person’s identity. There is a chance of fraud in the case of physical exams. The AI proctor system will also have a mechanism to detect the person’s voice, whether he is saying something or not. If the students are cheating, their behavior can be observed, and marks can be given accordingly. With less labor cost, this will be a tremendous technological advancement with higher accuracy, reliability, and adaptive than traditional conduction of exams. This system is built to reduce human error and save time. AI proctoring system will cut short the human error, and the detection of any suspicious activity is more likely to be caught by this rather than a physical teacher. Also, in this health crisis, where the need of the hour is to conduct exams remotely, AI proctoring system is the most viable and reliable solution. This system also allows the institutions to benefit from removing AI proctoring and disallowing students to engage or indulge in unfair means to pass the exams. Also, the online proctoring through AI can automatically scan and verify the student’s identity, and if it doesn’t match, he/she will not be eligible to give the exam. The system can also monitor competitive exams for those students who need to travel across cities for exams. This system can be helpful in these situations.

AI proctoring system can be a boon to the education system at a minimal cost. During exams, we hear about cheating news on media. Using AI proctoring system, these will be reduced. Also, in this health
7 POTENTIAL USE CASES OF ALGORITHMIC GOVERNMENT

crisis, where the need of the hour to conduct exams remotely, AI proctoring system is the most viable and reliable source. AI proctoring in the education department will play a significant role shortly, and it can remove the traditional approach of conducting exams. Due to any situation, the students cannot reach their destined locations for exams can give their exams online remotely. This AI proctoring system later can replace competitive exams where students need to travel cities to pass exams using the AI proctoring system; they can easily give exams sitting at their home. So, AI proctoring system will prove to be significant technological advancement in the education system.

7.4 ENTRANCE EXAMS AND ADMISSIONS

Problem of Admissions to Educational Institutions

University education has become a necessary part of people’s preparation for working life. Admission to the university is an important issue. How the student must choose the university and how the university must select a student on so many applicants. The success of both sides is determined through education. The students’ enrollment has increased in past years, which leads to more applications, more paperwork, and processing challenges. Every applicant’s forms are routed through different departments for evaluation and manual processing, which causes difficulty in the admission process. In today’s time, despite technological advancement still, the admission process at college or school admission is being carried out manually, which is very time-consuming. An automated student admission system’s development is the best solution to speed up and simplify the admission process and remove manual processing. The development of science and technology has immensely contributed to the growth of the internet, which has increased the need to develop an automated student admission process. The problems associated with the current system are that they publish multiple merit lists, leading to double work for the university executive and even for applicants to visit again to check the list. The manual admission process leads to the high cost of the application process because the colleges’ paperwork and admission fees are high. Even for the institute, they have to do all the manual work of handling the papers. With the advancement in technology, the process of admission can be done in an automated way by taking the data from previous year students, which are given admission to train a model that
can predict new applicants’ selection possibility. It will be beneficial for both the applicants and university admission cell to provide a better deserving candidate admission into the university program. This model will increase the process and accuracy to select the applicant. Before the entrance, every university considers a general aptitude; it can be GMAT, GRE, CAT, GATE, and others, but all these exams are conducted at centers to have a check on the identity of the person who has appeared for the exam. There are problems associated with this procedure. Every year millions of students give entrance test which leads to rush at the exam center and even on the roads many times leads to traffic jam because of this, students sometimes reach late, and they can’t appear for the test and miss the opportunity. To make this process completely automated, we can conduct the exams at home for students.

Requirement of the Online Entrance Process

In today’s world, the number of students who go for higher education has gradually increased. Based on data analysis from the UNESCO Institute of Statistics 2016, more than 300,000 Indian students have been gone out for higher studies. There is an annual growth rate of 22% in 16 years. The shift in trend for higher education has led to so many applications every year to the universities, making it difficult for them to decide which requests should be granted admission. This has become a major problem nowadays and till now the decision is taken by the admission cell on the manual basis which is time-consuming. Let’s look at the traditional admission method that was informing the students and applicants was done by putting up the notices on notice board and advertisements in the newspaper. The application form was a hard copy style where physically you have to go to the university to submit, and then evaluation is being carried out by the staff. With advancements in technology, the application form gets turned into a digital way, but the assessment is still done manually. The problem statement has brought a new change from paperwork to digital format, standing in a queue, and submitting the application form to submit the form online with just one click of a button. There is a need for a shift in the admission process for graduate education. Technological advancement can help to solve the problem of manual evaluation of each applicant’s form. According to the report Artificial Intelligence Market in the US Education Sector, AI will grow at a compound annual rate of 47.7 percent from 2018 to 2022. AI has the potential to change
how the admission process of domestic and international students can be changed by creating algorithms and models to predict the most likely applicant to be selected. Machine learning, the subset of AI, helps the computers analyze and learn from the university data, which factors are essential for admission. According to the Enrolment Management Report, AI can change the entry for small and prominent universities in both the public and private sectors.

**Algorithmic Solution**

With the advancement in technology, now, we can automate the whole process of admission through machines. From the entrance exam conduction to the result generation, document verification, and admission selection, all can be done with a machine’s help. The solution which is used to automate the entrance exam process is that while the applicant submits his identity card details, we can use face verification to check the identity of the person who is sitting in front of the computer to check that same applicant is giving the exam, or there is some cheating happening so we can easily terminate the exam of such faulty applicants. As shown in Fig. 7.5, the block diagram depicts the applicant’s real-time face verification at the time of giving the entrance test. The prototype code is shown in Appendix D.

After the entrance exam, the selected applicants apply for the document submission process. In the letter of recommendation, we can use sentimental analysis using natural language processing techniques to quickly get to know about positive or negative results about the person based on the advice given in the letter. The block diagram, as shown in Fig. 7.6, depicts the working of sentimental analysis.

After passing through the document verification stage, we can feed the data to the system to evaluate that the applicant has the chance of getting selected or not. Admission officers have a rich dataset of information of the past applicants, including the decision. Using this data, we employ supervised machine learning techniques to classify applicants.

For the classification problem, the information which has been taken is shown in Table 7.1. The dataset attributes will be pre-processed, and the missing values and outliers will be checked. After that, we will pass on the dataset to the classification models to train the model, and testing data will help verify the accuracy of the model. The classification model block diagram is shown in Fig. 7.7.
Analytical Techniques Used for Solution

Artificial Neural Network: Artificial neural networks (ANN) mimic the human body system’s biological brain. They are modeled after biological neurons, learn by the training given to them. The structure of ANN has three components input layer, hidden layer, and output layer. We use activation functions to convert the weighted sum of a neuron’s input signals into the output signal. In ANN, we use two methods, feedforward, and backpropagation, to tune the desired output model. In feed-forward, the input layer sends the input values to the hidden layer, which processes the signals and sends them to the output layer through an activation function. In the backpropagation method of supervised learning, the difference between desired and output result is measured. Then, the weights’ values are modified to obtain the nearest possible output value compared to the desired result.

Figure 7.8 shows the structure of ANN where we have taken all the variable from dataset, i.e., age, board, 12th%, GRE score, CGPA, and others as input variables. They are then passed on to the hidden layer using the sigmoid activation function, which is used in a binary classification where small change occurs in x, resulting in a large change in y. The
7 POTENTIAL USE CASES OF ALGORITHMIC GOVERNMENT

Collection of data

Data filtering and pre-processing
- Text conversion to lowercase
- Removing punctuations
- Removing stop words

Sentiment analyser classifier

Evaluation stage

Test data

Training data

Fig. 7.6 Sentiment analysis block diagram (Source Author creation)
Table 7.1  Dataset attributes considered for Prototyping

| Variable name | Description | Category |
|---------------|-------------|----------|
| Age           | Age of applicant | Quantitative |
| Board         | CBSE/ISCE/State | Categorical |
| 12th %        | Percentage in 12th class | Quantitative |
| GRE score     | GRE marks | Quantitative |
| Category      | Gen/OBC/ST/SC | Categorical |
| SOP           | Statement of purpose | Quantitative |
| LOR           | Letter of recommendation | Quantitative |
| CGPA          | Undergraduate CGPA | Quantitative |
| Research      | Research experience | Quantitative |
| Chances of admit | Chances of admission (0–1) | Quantitative |

The function is defined as shown in Eq. 7.3

\[ S(x) = \frac{1}{1 + e^{(-x)}} \]  

After the hidden layer processing, the result is sent to the output layer.

Logistic Regression: Logistic regression is a popular technique used in classification problems because it results in a value between 0 and 1. On the other hand, linear regression results can be positive or negative in cost, which cannot distinguish between an applicant’s application for higher education to be accepted or rejected, so we use logistic regression as the first method in most cases. The admission decision function is as follows.

Admission prediction = \( f(GRE \text{ score}, Research \text{ experience}, 12th \% \text{, and other variables}) \)

Decision Tree: A decision tree is a supervised machine learning algorithm used for classification problems. The decision tree comprises several branches, root nodes, and leaf nodes. It generated a tree-like structure by splitting the data according to specified parameters. Feature selection is based on the most significant information gain of features, and then the
iterative process repeats itself until it reaches the leaves. In our case, we have a feature like age, GRE score, 12th percentage, LOR, and other variables according to which the decision tree is formed. The root node contains the element with the most significant information gain of them
According to the problem statement, let’s see an example of a decision tree, as shown in Fig. 7.9.

In Fig. 7.9, an example illustrating the decision tree for a candidate’s admission prediction is shown. If a candidate has a CGPA of above 9.0, and either has done some research work or is having a GRE score of more than 320, has more chances of getting selected for admission to the university.

Random Forest: Random forest classification belongs to a supervised learning algorithm, an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the categories or means prediction of the individual trees. It smoothes the error which exists in a single tree,
and it increases the overall performance and accuracy of the model and provides explainable scores. The random forest provides us with an intuitive way to look at our features by listing individual feature importance, which gives us the importance of each factor affecting the loan approval. As compared with the decision tree, random forest works very fast, accurate, and robust. It can be used for both classifications and the regression problem.

**Impact of Algorithmic Solution**

The automated entrance exams and admission process will positively affect the system because the current system of manual exams and admission process is very hectic, time-consuming, and includes paperwork. With the automation of the whole process from entrance exams to admission, we can eliminate all the current system’s disadvantages. The solution will change the entire process of admission process in the higher education system. In the entrance exams, instead of going to the exam centers, we can automate the exams, and students can give them while sitting at home. It will save the applicant’s travel cost who will be traveling from a remote area to appear for the exam at the center. We can make a big bank of question pool through this process so that there will be no chances of cheating. With the microphone and webcam on, we can keep a check on the applicant. The face verification model will continuously track the
applicants’ face to know that there is no cheating happening while giving the exam. With the auto-grade procedure, we will save the applicant and the university’s time because the result will be given to the applicants within few hours, and then they can move to the next stage of the admission process. Even in the job aptitude exams, we can use these automated entrance exams, and the applicant can easily give it from their home.

At the time of document identification and submission of papers, we can easily use the natural language processing model to evaluate them. In the letter of recommendation (LOR), which is submitted by the applicant for higher education, we can assess the message by sentimental analysis technique. We can rate the student’s feedback as positive or negative, which is considered an essential factor at the time of evaluation of an applicant because it tells about the applicant’s personality and past work history. With such an admission system, the applicant can have a real-time check on the status of his/her application to a university. After all, the data from documents are verified and collected is stored in the database.

With the help of supervised classification learning, we can evaluate the applicant’s data. This will help the admission officers quickly and accurately judge a student’s application without any manual evaluation done by a human. The classification model will easily predict the chances of an individual in the university. This will make the workload of the admission cell lighter and more accurate. This automated system will create a positive impression on students and engage the students with clear and transparent information in students’ selection. It will help to select the best talent from the pool of applicants. Students prefer to interact via the website and mobile devices with the hyper-growth of digital technologies, which increases student satisfaction.

7.5 Funding & Loans

The Problem of Loan Frauds

In the banking industry, the distribution of loans is the core business of every bank. The central bank’s asset directly came from the profit earned from loan distribution by the bank. The prime objective in the banking industry is to invest its assets into safe hands. Banks and financial companies approve a loan after the validation and verification process, but still, there is no guarantee that the selected applicant is a deserving applicant or not. Due to insufficient credit history, many people face loan rejection.
These people are students or unemployed adults because they do not have enough knowledge to justify their credibility. For example, an employed adult has an income source, which turns out to be a significant factor in repayment of the loan. There are many factors like real estate, marriage, city of residence, and others, which play a credit score history for loan approval.

In today’s world, the banks struggle to gain the upper hand over each other to enhance their business. Retaining customers and fraud detection are the two critical goals in the banks. There are many risks for the bank, as well as those who get the loan. Stake in the bank involves credit risk, i.e., the loan won’t be returned in time or at all, the interest rate may be too low that the bank won’t be able to earn adequate money. Risk management is widely used in the banking sector, and the executives need to know the credibility of the customer they are dealing with. Offering the customer’s credit card or loan is always a risky job.

With the advancement in technology, the banks focus on the automation process of fraud detection in the loan approval or credit card. Manual processing of the loan lacks consistency and accuracy; above all, it is time-consuming. Automation in loan approval processing facilitates faster and more accurate loan applications, which make the whole process seem less dreadful, faster, and more reliable. More rapid loan processing is always a competitive advantage for them. The availability of a vast amount of data helps the banks to enhance their loan lending operation by implementing the loan lending prediction. The bank’s historic data plays a crucial role in training the model to predict loan approval. The loan prediction is helpful for the banks and applicants because the main goal is to provide the loan to the right applicant as fast as possible. The loan prediction system helps the executives directly jump to a specific application without wasting time on other apps, and it can be checked on a priority basis. The close monitoring of why one applicant’s loan was approved and others were rejected such data is beneficial to the customers and the banks.

**Importance of Detecting Fraudulent**

Credit scoring is a process of evaluation before the credit score is sanctioned. This process is called credit evaluation, which concludes the approval or rejection. Credit scoring plays a vital role in evaluating the applicant’s loan approval or denial in the banking sector. There are 5C’s involved in the process in credit evaluation, which are character, credit
A person who sounds financially stable is likely granted a loan very quickly. Credit history is another critical feature to approve the credit application. The credit report includes the person’s past transactions, borrowing, and other functions. If the person has a good cash flow, then the chances of getting credit approval are much higher.

As per a study, more than 50% of the first-time loan applicants get rejected from the banks and financial institutions because the existing predicting system works on credit scores and paperwork. Due to the high demand and the reliability of the banks and financial institutions on loan lending, there is a demand for further improvement of the credit scoring model. As per the study conducted by the National Business Research Institute and Narrative Science, around 32% of financial service providers in the country have already begun using AI tech. As the world is changing, the model for evaluating the credit score has to change. In India, 80% of the population doesn’t have a credit score, which leads to the rejection of the loan approval requests. If, with the help of Artificial Intelligence, we can build creditworthiness, it will open a big business portal for the banks and financial institutes. Nowadays, fintech companies are taking the data from the person’s transaction, like how much he/she spends on food, travel, clothing, and more. All this data can be used to create a person’s creditworthiness rather than relying on old factors like credit history and credit score.

The importance of the existing problem in the banking sector helps find new fields in the world to make the current system more robust and refined. As discussed earlier new system will open large business portals for the banking sector. From the expert’s point of view, the AI can help lenders, banks, and financial institutions to reach over 300 million new first-time loan applicants, which will boost the economy. According to a report by the Boston Consulting Group, the use of AI in loan assessment can help digital lending to grow to USD 100 Billion business by 2023 in India.

**Algorithmic Approach**

The above problem statement can be solved using the machine learning methodology. We can use the bank’s historical data from the loan applications and all the documents attached to the application for loan approval. The process of loan evaluation is a sequence of steps taken to grant
the loan to the right applicant. The proposed model for the problem statement is shown in Fig. 7.10.

Business Understanding: The initial phase focuses on understanding the project objective and the required business perspective forms. In the banking sector, we need knowledge about the loan approval process, essential documents required, criteria on which decision is made for the approval and rejection.

Data Gathering & Understanding: The process of gathering the data depends on the problem statement we can choose for the real-time data or collecting the data from various files or databases. There are free datasets available on the website, like Kaggle and machine learning repositories. The data understanding phase focuses on familiarizing data, identifying the relation of features with the target value, and getting to know more about the data. Generally, we apply an 80:20 split to the data for the training and testing process, respectively. A collection of data from the bank sector has been selected for the prototype shown in Appendix E. Table 7.2 indicates the name of attributes, description, and category which are used to train the model.

Data Filtering: The bank dataset attributes are filtered, and the relevant attributes needed for prediction are selected. The process of cleaning the raw data is termed data preprocessing. The handling of missing values, outliers, and human errors is corrected before passing the model’s inputs. The preprocessing is done because noisy data leads to inconsistency. Handling noisy data improves the efficiency of the algorithm.
### Table 7.2  Dataset attributes for the Prototyping

| Variable name               | Description                      | Category       |
|-----------------------------|----------------------------------|----------------|
| Loan ID                     | Unique loan ID                   | Qualitative    |
| Age                         | Age of the applicant             | Qualitative    |
| Gender                      | Male/Female                      | Categorical    |
| Married                     | Marital status                   | Categorical    |
| Dependents                  | Number of dependents             | Qualitative    |
| Education qualification     | Graduate/Not graduate            | Categorical    |
| Housing                     | Own/Free/Rent                    | Categorical    |
| Self employed               | Yes/No                           | Categorical    |
| Credit card                 | Applicant credit card (0/1)      | Qualitative    |
| Existing client             | Applicant is an existing client (0/1) | Qualitative    |
| Applicant income            | Income                           | Qualitative    |
| Co-applicant income         | Income                           | Qualitative    |
| Payment delay               | Is there any payment delay (0/1) | Qualitative    |
| Average house spend         | Household spend monthly          | Qualitative    |
| Existing loan               | Is there any existing loan (0/1)  | Qualitative    |
| Average travel              | Traveling expense                | Qualitative    |
| Present account balance     | Account balance                  | Qualitative    |
| Minimum balance             | Minimum balance in 6 months      | Qualitative    |
| Loan amount                 | Amount                           | Qualitative    |
| Loan amount term            | Term of the loan in months       | Qualitative    |
| Credit history              | credit history meets guidelines  | Qualitative    |
| Property area               | Urban/Semi-urban/Rural           | Categorical    |
| Loan status                 | Loan approved (Y/N)              | Categorical    |

**Searching Model:** Our main goal is to train the best-performing model using the filtered data. In our case, we have the target variable, that applicant should get loan approval or rejection based on the model’s input features. We will use the classification problem because we want the output to be true or false, yes or no. The most commonly used classification algorithms are K-nearest neighbor, logistic regression, decision tree, and random forest.

**System Evaluation:** The best-performing models are evaluated during the model development process, finding the best-performing model on the filtered data. Then we apply that model to the test data to check the performance.

The block diagram for the solution is shown in Fig. 7.11.
Problem Statement

Identify the required data

Data collection

Exploratory Data Analysis

- Handling Missing Values
- Outlier detection and removal
- Label encoding categorical data

Data Pre-Processing

Visualizing the categorical data and Numerical data with the target variable.

SPLIT

Training Data

XGBoost Feature Importance

Feature selection

- Random Forest
- Logistic Regression
- XGBoost classifier
- Decision Tree

Compare Accuracy

Evaluation of test data

Test Data

In dataset, there are 13 existing features
10 addition feature has been added

Fig. 7.11 Block diagram for loans & funding (Source Author creation)
Techniques Used for Algorithmic Solution

Logistic Regression: Logistic regression technique is the most popular statistical technique in the financial sector world, especially for the credit risk assessment or loan approval prediction. As compared to linear regression, it overcomes multiple issues, such as in linear regression, and we can get a negative or positive output, which is not possible in probability. In contrast, logistic regression provides a continuous range of values between 0 and 1. We can assume that the likelihood of loan approval follows the logistic distribution; it follows the function of loan approval, which is written as follows.

\[
\text{Loan approval} = f(\text{credit history, applicant Income, Education qualification, Employment status, and other dummy variables})
\]

Where the Loan approval decision is 1, and the rejection is 0.

In logistic regression, to map the predicted values to probability, the sigmoid function is used to map them. The sigmoid function is shown in Eq. 7.4

\[
f(x) = \frac{1}{1 + e^{-x}}
\]  

(7.4)

In the case of a classification problem, it is a great model to try as the first step because it outputs conditional probability value; however, we find predicted value in regression.

Decision Tree: A decision tree is a supervised machine learning algorithm used for classification problems. The banking industry is widely used for credit risk assessment as the decision tree comprises several branches, root nodes, and leaf nodes. It generated a tree-like structure by splitting the data according to specific parameters. Feature selection is based on the most significant information gain of features, and then the iterative process repeats itself until it reaches the leaves. An example of a decision tree is depicted in Fig. 7.12 related to the banks and financial institutions’ loan approval process.

In the decision tree, there is a set of rules which are trained using the training data. It ignores the irrelevant features which are not required. Figure 7.12 shows that if an applicant who applies for the loan application has an excellent credit rating gets a loan approval very quickly. The
A decision tree model is suitable for small and straightforward datasets as it is easy to interpret the tree. Still, when the data gets more complicated, its accuracy decreases and gets challenging to interpret.

Random Forest: Random forest classification belongs to a supervised learning algorithm, an ensemble learning method for classification that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the categories or means prediction of the individual trees. It smoothes the error which exists in a single tree, and it increases the overall performance and accuracy of the model and provides explainable scores. The random forest provides us with an intuitive way to look at our features by listing individual feature importance, which gives us the importance of each factor affecting the loan approval. As compared with the decision tree, random forest works very fast, accurate, and robust, and it can work on large datasets and deal with the unbalanced dataset. It can be used for both classifications and the regression problem.

XGBoost: XGBoost stands for extreme gradient boosting is a scalable and highly efficient boosting system. It is an ensemble of decision trees algorithm where trees are added until no further improvements can be made to the model. The benefit of using XGBoost is that it helps determine the critical feature in the dataset and helps to evaluate the model on them for better results. It provides a score that indicates how useful a feature is in the construction of the model. In the solution, we have used XGBoost for feature selection and classification.
Impact of Algorithmic Solution

Automation in the loan approval processing has distinct advantages over the existing system. In a week, banks and financial institutions received thousands of loan applications, and only a few percentages of them are approved. But with automated loan approval processing gives the ability to quickly evaluate the thousands of applications and find the right applicants that meet the credibility criteria. Replacing the manual steps with automation and transferring the paperwork to a digital platform leads to a better user experience and improve the speed and accuracy of loan approval. As per the Federal Reserve survey in 2016, half of the applicants complained about the difficulty during the application process and the time taken to receive the credit decision. An automated lending system can dramatically change this perception by simplifying and accelerating the credit decision’s entire operation. The digitalization of documents will make automation in the loan approval process quick because all the required information will be collected in no time and transferred to the system for the decision evaluation.

In today’s auto credit lending market, a faster credit decision is a competitive advantage for both banks and financial institutions. Suppose a small bank has a much quick and automated decision system for loan approval. In that case, the applicants applying for the loan are much more satisfied, which helps the bank retain its customer, which is the bank’s main objective, and the bank provided stiff competition to the others.

Automation in the banking sector will optimize all the operations. Most of the banks already have automated banking services, and automating the core services like loan approval will be another step in the industry. All the operations will be done online, producing more clear data, and enhancing the loan approval process model. Even the loan process’s management visibility can easily be seen by the applicant, which makes it fairer to know the rejection of the application.

The automation process will reduce loan processing cost as the bank companies do not have to spend the money on training the employees for service loans. Even the automation process will save the time of the executive as time is money in every industry. The decrease in the paperwork will be an added advantage to companies’ cost reduction in handling the paperwork, manually entering the customers’ financial information to remove the paperwork from the process of credit lending.
7.6 Online Reputation Management

The Problem of Online Reviews for Government

Many Governments across the world are trying to move closer to their citizens to achieve transparency and engagement. In recent years, there has been growing interest in mining the online sentiment to predict how a decision made by the Government is taken by the public, either positively or negatively. The explosion of social media is opening new opportunities to help the Government achieve it. People are continuously using social media daily to communicate and express their opinions about various subjects, products, and services. It has emerged as a precious resource for text mining and sentiment analysis. Social media communications include Facebook, Twitter, etc. Twitter is one of the most widely used social media sites used by citizens. In this use case, an attempt has been made to analyze the citizen sentiment in social media such as Twitter. Primary data was collected through Twitter API. An approach to a real-world problem and how Government agencies can benefit from it as described in this section.

A measure of the customer’s attitude toward the aspects of a service or product which they describe in text or a computational study of opinions and emotions is called the sentiment analysis. Sentiment analysis is in demand nowadays because of its efficiency and ability to analyze even huge texts in minutes. The task of analyzing the sentiments typically involves taking a piece of writing, which can be either a sentence, a comment, or an entire document, and returning a ‘score,’ which is a measure of how positive, negative, or neutral the text is. This helps evaluate the Government’s performance from the people’s perspective instead of making people’s surveys, which might be expensive and time-consuming. Sentiment analysis is also known as opinion mining, and subjectivity analysis determines the attitude or polarity of opinions given by humans for a particular scheme. Polarity is the quantification of sentiments with a negative or positive value. Sentiment analysis can be applied to any textual form of opinions such as reviews, blogs, and Microblogs. Microblogs are small text messages such as tweets or a short message with a limitation of 160 characters. In previous years, Twitter has gained more and more popularity and is used as a microblogging website. The messages in Twitter, or tweets, are a way to publicly share and express a defined group of users’ interests. With the limitation of the tweets’ size, it has challenged the users to express their emotions in one or two key
lines or sentences, and hence, it gives a fair reflection of what is happening across the country and the world.

The complete process of identifying and extracting the subjective data from the available raw data is called sentiment analysis. Natural language processing, which describes the relationship between human language and machine, tries to narrow down the gap between humans and machines by extracting useful information from the natural language messages. In this case-study, the extraction of the sentiments from the tweets is studied.

**Importance of Online Sentiments**

Sentiment analysis is a technique to discover knowledge with the help of data mining. Also known as opinion mining, this can be applied to unstructured data sources and can cover various information on various topics related to politics and government. According to an estimate, 80% of the data is unstructured or unorganized. A considerable volume of data is in documents, emails, media conversations articles, and surveys created every day, which is very hard to analyze computationally very time-consuming. Sentiment analysis helps by tagging the unstructured data and sorting data. Sentiment analysis is considered a beneficial approach in predicting the current trends, interpreting the nation’s reaction and nerve. Opinions of people can be evaluated, and suggestions can be implemented. The workflow of governance with sentiment analysis can result in better and efficient use of power and energy in the correct direction.

Sentiment analysis can identify critical issues and that too in real-time. Sentiment analysis helps government sites identify the reactions and sentiments of people on a program or policy created by them for the welfare of the citizens. Different government sites can analyze the response of the public on their strategy and work for the betterment. In 2013 US presidential elections sentiment analysis was used. Governments can use the data from the social media platforms to analyze for predicting and the opinion of the common public. News media has always been using sentiment analysis to predict trends. The project of the government, whether it was successful or not, is analyzed using sentiment analysis. Opinion polls of the Indian elections of 2019 were predicted using major news media platforms’ sentiment analysis. Sentiment analysis on the above problem helps the governments prioritize work per the demands of the citizens and channelize the workflow of the government that puts the needs and services of the citizens rather than the governments themselves.
Extracting data and analyzing the reactions and based on the responses, the policies can be modified and created, keeping the citizens’ view and opinion first.

Sentiment analysis is in trend already and is being used by most governments around the world to build a citizen-centric governance model. It is sensed to be quite a popular tool for the governments to analyze their policies and the workflow in which they work. In the coming decade, governments will have a tremendous amount of involvement from its citizens. Governments and citizens will work together in sync with the help of sentiment analysis. Governments can set up a criterion for its policy to pass or fail based on the review of the country’s people. The new welfare programs for the citizens and what they require next can be analyzed, and based on that, knowledge gained can bring up new plans that cover up the needs and necessities of the people around the country. Governments will get a clear view of how people look forward to their policies and programs.

Algorithmic Solution

The approach used for solving the sentiment analysis problem that has been adopted for the prototype shown in Appendix F is presented in Fig. 7.13 and described below.

1. Data collection: The first step of sentiment analysis includes collecting data from blogs, social networks, and forums. The data is first scrapped from the official Government site. Based on the scraped information, the tweets from Twitter are extracted and saved into a CSV file. The data collected is unorganized and in different languages; therefore, text classification and natural language processing are used to extract and classify the data.

2. Data Preprocessing: The data preprocessing step comprises cleaning the noisy and incomplete data. The irrelevant and non-textual

![Fig. 7.13 Sentiment analysis approach (Source Author creation)](source)
content are identified and eliminated before analyzing the data. The preprocessing for the sentiment analysis includes the following tasks.

- Removing the URLs, punctuation, special characters, etc.
- Removing the Stopwords.
- Removing Retweets.
- Stemming.
- Tokenization.

3. Sentiment detection: Sentiment detection is a fundamental work in various sentiment analysis and opinion mining applications, such as tweets mining and tweet classification. The extracted sentences of the reviews and opinions are examined, and the sentences with subjective expressions such as opinions, beliefs, or views are retained. Sentences with external communication like facts or factual information are discarded. The words for the sentiment words can be classified into positive, negative, and neutral words.

4. Sentiment polarity: TextBlob is a Python library for processing textual data. TextBlob provides an API for diving into standard natural language processing (NLP) tasks such as part-of-speech tagging, sentiment analysis, classification, translation. Here, TextBlob is used for calculating the polarity as well as classifying into sentiments. Polarity is a float value within the range $[-1.0 \text{ to } 1.0]$ where 0 depicts neutral, +1 depicts a very positive sentiment, and −1 represents a negative sentiment. Subjectivity is a float value within the range [0.0 to 1.0] where 0.0 is hugely objective, and 1.0 is exceptionally subjective. The subjective sentence expresses some personal feelings, views, beliefs, opinions, allegations, desires, and beliefs, whereas objective sentences are factual.

5. Sentiment classification: based on the previous step’s polarity, subjective sentences are classified in positive, negative, good, bad, like-dislike, but classification can be made by using multiple points. The sentiments derived from the sentences can be divided into three categories, as shown in Fig. 7.14.

6. Analysis of tweets: The main objective of performing sentiment analysis is to convert unstructured text into meaningful information. So, the analysis of results is vital to decide by the government. In the case of government schemes announced by the central government, if more tweets result is positive, then people support that particular scheme. Analysis can be used to take feedback on specific plans with
appropriate discussions in public and decide for proper implementation of government schemes. This can also be visually plotted on graphs like a pie chart, bar chart, and line graphs. When the analysis is finished, the results are displayed with the percentage of positive, negative, or neutral tweets.

**Analytical Techniques for Sentiments**

The solution to this problem requires the application of a few of the following analytical techniques.

1. Web Scraping: Web scraping is also known as web data extraction. It refers to the process of retrieving or ‘scraping’ data from a website. Unlike the regular operation of manually extracting data, web scraping makes use of intelligent automation to retrieve data points from the internet. It is a technique used to pull large amounts of data from websites whereby the data is extracted and saved to a local file on a computer or PC. Web scraping can be used for sentiment analysis to obtain the data from any Government site or organization. First, the web scraper will be given one or more URLs to load the page before scraping. The scraper would then load the entire HTML code for the page. In this project, web scraping has been used to extract the data from the official Government website, mygov.in. The scraper would either remove all the data on the page. The views of the people are obtained from the discussion forum on
a particular topic. Lastly, the scraper would output all the data it has extracted from the website.

2. Natural Language Processing and Information extraction: After web scraping the data from the Government site, natural language processing was applied to get the most frequent words and extract the Twitter data related to the most frequently occurring word. The information was then stored in a CSV file. Natural language processing libraries were used for the preprocessing of the text. The data collected from the government websites was used to find out the frequent word discussed and accordingly found the tweets for Twitter and analyzing those tweets for sentiment analysis. The tweets selected were preprocessed using a natural language processing tool kit, and different sentiments were analyzed. We used the natural language processing technique as this is best suited for the preprocessing of the text data and was feasible for our problem statement.

3. Text analysis and classification: After applying the Natural language processing techniques and preprocessing the data, removing the URLs, hashtags, usernames, lemmatization, stemming, etc. TextBlob was used to calculate the polarity of sentiments and classify them into positive (if the polarity is greater than 0), negative (if the polarity is less than 0), or neutral views. TextBlob is based upon Pattern Analyzer and Naïve Bayes Analyzer. The columns for the calculated polarity and the classification of sentiments were added. For classification of the tweets and labeling them as positive and negative is done using classification techniques. TextBlob has an inbuilt library for classifying sentiments using the subjectivity and polarity of the text. Using that sentiment analysis of TextBlob, it is much easier to scale tweets and classify them according to the problem statement.

4. Visualizing the results: After the classification of sentiments, the percentage of the total number of positive, negative, or neutral tweets is calculated individually. These tweets can be visualized easily using matplotlib and pyplot libraries. Visualizing the results is much easy to analyze and depict what is happening in the entire code.
Impact of the Algorithmic Solution

In today’s world, governments are moving toward citizen-centric governance where the priorities are services driven by the citizens instead of the government. Today analyzing and getting into the nerves of common people is the new form of governance. Every government wants to know the impact or the reaction of the citizens on any new policy or act or decision. Governments today wish to take the people’s opinions and get to the real perspective of the country’s people so, if they could analyze the citizens’ reaction and further work more to improve the decision or law that could help a government stay in power and understand the real situation and work upon that.

This is the age of social media. Hashtags, mentions, and whatnot today’s generation express themselves quite well on social media, so the best form of opinions reactions and suggestions can be gathered from that public sourced data. Governments have realized that social media is an excellent vehicle for getting closer to the citizens. The solution above will have a massive impact on the way governance will change to citizen-centric administration. This model will help the government understand the citizens, how they think, and how they can make changes to better the country and society. The above solution will take the discussions from the government sites initiated by ordinary people. The most common word on the discussion form will be selected and searched for on the social media platform and will analyze accordingly how citizens feel about that policy, law, or decision made by the government. Sentiment analysis is a useful tool for the governments and can help determine whether a particular program initiated by them was successful or a complete failure. This solution’s impact is that this will create a bridge between the government and the citizens’ opinion. With the help of sentiment analysis, governments can reprioritize their policies and reallocate their funds on how citizens can gain from those funds. All this sentiment analysis activity can be done at a minimal cost while also getting the common people’s trends and reactions. Measuring the customer’s attitude toward the aspects of a service or product described in the text or a computational study of opinions and emotions can help the government. Governments can study the sentiments and accordingly modify or make changes in the policies and laws. This type of sentiment analysis on the public’s views can be used to segregate different sections of society. It can be used to predict the average percentage of supporters of the government. Harsh comments
can be directly eliminated to maintain the balance between the sentiments and get accurate results. The repeated Retweets by a person can be removed so that accuracy is calculated based on all the persons’ views and not affected very much by a single person.

Although these are the early stages of this in the future with the amount of data increasing and getting such a massive volume of opinions and suggestions can make the governance more robust and flexible at the same time. Citizens never were so crucial in the country’s progress, but this will help in betterment due to the sentiment analysis. People can openly express and judge the government’s initiatives, and the government can make the necessary changes required as per the demands of the people.

**Further Readings**

Alessia, D., Ferri, F., Grifoni, P., & Guzzo, T. (2015). Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications, 125*(3).

Alessio, H. M., Malay, N., Maurer, K., Bailer, A. J., & Rubin, B. (2017). Examining the effect of proctoring on online test scores. *Online Learning, 21*(1), 146–161.

Arun, K., Ishan, G., & Sandmeet, K. (2016). Loan approval prediction based on machine learning approach. *IOSR Journal of Computer Engineering, 18*(3), 18–21.

Arunachalam, R., & Sarkar, S. (2013, October). The new eye of government: Citizen sentiment analysis in social media. In *Proceedings of the IJCNLP 2013 workshop on natural language processing for social media (SocialNLP)* (pp. 23–28).

Aslam, U., Tariq Aziz, H. I., Sohail, A., & Batcha, N. K. (2019). An empirical study on loan default prediction models. *Journal of Computational and Theoretical Nanoscience, 16*(8), 3483–3488.

Bakliwal, A., Foster, J., van der Puil, J., O’Brien, R., Tounsi, L., & Hughes, M. (2013, June). *Sentiment analysis of political tweets: Towards an accurate classifier*. Association for Computational Linguistics.

Basu, B., & Srinivas, V. V. (2016). Regional flood frequency analysis using entropy-based clustering approach. *Journal of Hydrologic Engineering, 21*(8), 04016020.

Bramer, M. (Ed.). (2006). *Artificial intelligence in theory and practice: IFIP 19th World Computer Congress, TC 12: IFIP AI 2006 Stream, August 21–24, 2006, Santiago, Chile* (Vol. 217). Springer Science & Business Media.
Bryant, P., & Ruello, J. (2019). One system to examine them all: Defining the complexities of implementing an institution wide online exam model. *Personalised Learning. Diverse Goals. One Heart.*, 370.

Corallo, A., Fortunato, L., Matera, M., Alessi, M., Camillò, A., Chetta, V., ..., Storelli, D. (2015, July). Sentiment analysis for government: An optimized approach. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 98–112). Cham: Springer.

Coșer, A., Maer-matei, M. M., & Albu, C. (2019). Predictive models for loan default risk assessment. *Economic Computation & Economic Cybernetics Studies & Research, 53*(2).

Fong, S., & Biuk-Aghai, R. P. (2009, December). An automated university admission recommender system for secondary school students. In *The 6th International Conference on Information Technology and Applications* (p. 42).

Gao, X., Wen, J., & Zhang, C. (2019). An improved random forest algorithm for predicting employee turnover. *Mathematical Problems in Engineering, 2019*.

Goyal, A., & Kaur, R. (2016). A survey on ensemble model for loan prediction. *International Journal of Engineering Trends and Applications (IJETAB), 3*(1), 32–37.

Gupta, M., Iyer, K. K., Singh, M. R., & Kadam, A. K. (2017). Automated online college admission management system. *International Journal of Computer Science Trends and Technology, 5*(3).

Hamid, A. J., & Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal (MLAIJ), 3*(1).

Hubert, R. B., Estevez, E., Maguitman, A., & Janowski, T. (2018, May). Examining government-citizen interactions on Twitter using visual and sentiment analysis. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age* (pp. 1–10).

Kaur, A., & Hasija, S. (2015). A conceptual model of admission system and performance evaluation for a university. *International Journal of Computer Applications, 125*(4).

Li, M., Sikdar, S., Xia, L., & Wang, G. (2020). Anti-cheating online exams by minimizing the cheating gain.

Liang, Y., Jin, X., & Wang, Z. (2019). Loanliness: Predicting loan repayment ability by using machine learning methods.

Lux, T., Pittman, R., Shende, M., & Shende, A. (2016, May). Applications of supervised learning techniques on undergraduate admissions data. In *Proceedings of the ACM International Conference on Computing Frontiers* (pp. 412–417).

Madane, N. (2019). Loan prediction analysis using decision tree. *Journal of the Gujarat Research Society, 21*(4s), 214–221.
Mirza, M. M. Q., Dixit, A., & Nishat, A. (Eds.). (2003). *Flood problem and management in South Asia*. Kluwer Academic.
Purohit, S., & Kulkarni, A. (2011, December). Credit evaluation model of loan proposals for Indian Banks. In *2011 World Congress on Information and Communication Technologies* (pp. 868–873). IEEE.
Sankarasubramanian, B., Rathinakumar, M., Balakrishnan, B., & Bharathwajan, R. (2014). An automated implementation of Indian university admission system using artificial neural networks. *ICTACT Journal on Soft Computing, 4*(3), 767–771.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR), 34*(1), 1–47.
Sivasree, M. S., & Rekha Sunny, T. (2015). Loan credibility prediction system based on decision tree algorithm. *International Journal of Engineering Research & Technology (IJERT), 4*(9).
Syaidudin, Y. W., & Puspitasari, D. (2017). Twitter data mining for sentiment analysis on peoples feedback against government public policy. *MATTER: International Journal of Science and Technology, 3*(1).
Takahashi, M., Moro, A., Ji, Y., & Umeda, K. (2020). *Expandable YOLO: 3D object detection from RGB-D images*. arXiv preprint arXiv:2006.14837.
Vargiu, E., & Urru, M. (2013). Exploiting web scraping in a collaborative filtering-based approach to web advertising. *Journal of Artificial Intelligence Research, 2*(1), 44–54.
Walczak, S., & Sincich, T. (1999). A comparative analysis of regression and neural networks for university admissions. *Information Sciences, 119*(1–2), 1–20.
Waldrop, M. M. (2017). News feature: The genuine problem of fake news. *Proceedings of the National Academy of Sciences, 114*(48), 12631–12634.
Wang, M., Yu, J., & Ji, Z. (2018). *Credit fraud risk detection based on XGBoost-LR hybrid model*.
Waters, A., & Miikkulainen, R. (2014). Grade: Machine learning support for graduate admissions. *AI Magazine, 35*(1), 64.
Zhou, L., & Wang, H. (2012). Loan default prediction on large imbalanced data using random forests. *TEKOMNIKA Indonesian Journal of Electrical Engineering, 10*(6), 1519–1525.
Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science, 162*, 503–513.