**SENTIMENT ANALYSIS USING MACHINE LEARNING**

**Project Report**

MAJOR-PROJECT (ICI652)

Degree

**BACHELOR OF COMPUTER APPLICATION(CTIS)I-NURTURE**

|  |  |
| --- | --- |
| PROJECT GUIDE:  **Mr. VIJAY PRAKASH**  **SENIOR IT FACULTY** | SUBMITTED BY:  **TANYA PAL(TCA2056023)**  **MANISHA SINGH(TCA2056014)** |

APRIL,2023



**FACULTY OF ENGINEERING & COMPUTING SCIENCES**

**TEERTHANKER MAHAVEER UNIVERSITY, MORADABAD**

**DECLARATION**

We hereby declare that this Project Report title **SENTIMENT ANALYSIS USING MACHINE LEARNING** submitted by us and approved by our project guide, Faculty of Engineering & Computing Sciences. Teerthanker Mahaveer University, Moradabad, is a bonafide work undertaken by us and it is not submitted to any other University or Institution for the award of any degree diploma / certificate or published any time before.

|  |  |  |
| --- | --- | --- |
| **Student Name:** | TANYA PAL |  |
| **Student Name:** | MANISHA SINGH |  |

Table of Contents

[1 SENTIMENT ANALYSIS USING MACHINE LEARNING 3](#_Toc31139962)

ABSTRACT

1.1 DOMAIN

1.2 PROBLEM STATEMENT

2 PROJECT DESCRIPTION

2.1 Scope of the work…………………………………………………………………………………………………………………..4

2.2 Project Modules……………………………………………………………………………………………………………………..4

3 E-R DIAGRAM………………………………………………………………………………………………………………………………5

4 TECHNOLOGIES TO BE USED………………………………………………………………………………………………………..7

4.1 Software platform………………………………………………………………………………………………………………7

4.2 Hardware platform…………………………………………………………………………………………………………….7

5 Advantages of this Project…………………………………………………………………………………………………………….7

6 Future Scope and further enhancement of the Project…………………………………………………………………7

7 TEAM DETAILS……………………………………………………………………………………………………………………………..8

8 CONCLUSIONS…………………………………………………………………………………………………………………………….. 8

9 REFERENCES…………………………………………………………………………………………………………………………………9

**1.Project Title : SENTIMENT ANALYSIS USING MACHINE LEARNING**

**Abstract:**

Sentiment Analysis is a powerful technique that involves identifying and extracting subjective information from text data. It has gained tremendous popularity in recent years due to the explosion of social media platforms, which generate a vast amount of unstructured text data every day. In this project, we will use Machine Learning algorithms to perform Sentiment Analysis on airline reviews.

# 1.1Domain

Tool- JUPYTER NOTEBOOK

Technology- PYTHON, MACHINE LAERNING , NUMPY, PANDAS

# 1.2Problem Statement

The problem we are addressing in this project is to perform Sentiment Analysis on airline reviews using Machine Learning algorithms. With the growth of social media platforms, airlines are receiving feedback from customers in the form of tweets, reviews, and other social media posts. This feedback can be valuable to airlines to improve their services and understand their customers' needs. However, analyzing this feedback manually can be time-consuming and inefficient. Sentiment Analysis can automate this process and provide airlines with quick insights into their customers' sentiments.

Our goal in this project is to build a model that can accurately classify airline reviews into three categories: positive, negative, and neutral. We will use the Airline Twitter Sentiment dataset, which contains tweets about various airlines. The dataset has three classes: positive, negative, and neutral. The dataset is unbalanced, with a majority of tweets being negative.

# 2.Project Description

The objective of this project is to perform Sentiment Analysis on airline reviews using Machine Learning algorithms. Sentiment Analysis is a powerful technique that involves identifying and extracting subjective information from text data. It has gained tremendous popularity in recent years due to the explosion of social media platforms, which generate a vast amount of unstructured text data every day. In this project, we will use Sentiment Analysis to analyze airline reviews and understand customers' sentiments about the airlines.

Dataset: For this project, we will use the Airline Twitter Sentiment dataset, which contains tweets about various airlines. The dataset is available on Kaggle and can be downloaded from the following link: https://www.kaggle.com/crowdflower/twitter-airline-sentiment

The dataset has 14,640 tweets, and each tweet is labeled as positive, negative, or neutral. The dataset is unbalanced, with a majority of tweets being negative. The dataset contains the following columns:

The dataset has 14,640 tweets, and each tweet is labeled as positive, negative, or neutral. The dataset is unbalanced, with a majority of tweets being negative. The dataset contains the following columns:

tweet\_id: the ID of the tweet

airline\_sentiment: the sentiment of the tweet (positive, negative, or neutral)

airline\_sentiment\_confidence: the confidence level of the sentiment classification

negativereason: the reason for the negative sentiment (if applicable)

negativereason\_confidence: the confidence level of the negative reason classification (if applicable)

airline: the name of the airline

name: the Twitter handle of the user who tweeted

retweet\_count: the number of times the tweet was retweeted

text: the text of the tweet

tweet\_coord: the coordinates of the tweet (if available)

tweet\_created: the date and time when the tweet was created

tweet\_location: the location of the tweet (if available)

user\_timezone: the time zone of the user who tweeted

## Scope of the Work

The scope of this project includes the following tasks:

Data Collection: Collecting the Airline Twitter Sentiment dataset from Kaggle.

Data Preprocessing: Preprocessing the text data by removing mentions, URLs, special characters, punctuations, stopwords, and stemming.

Exploratory Data Analysis: Exploring the dataset to gain insights into the data, such as the distribution of sentiments, the frequency of words, and the most common reasons for negative sentiments.

Model Selection: Selecting the appropriate Machine Learning algorithm for Sentiment Analysis, such as Logistic Regression, Naive Bayes, Random Forest etc.

Feature Extraction: Extracting features from the preprocessed text data using CountVectorizer or TF-IDFVectorizer.

Model Training: Splitting the dataset into training and testing sets and training the model on the training set.

Model Evaluation: Evaluating the performance of the model on the testing set using metrics like accuracy, precision, recall, and F1-score.

Model Deployment: Deploying the trained model to classify new airline reviews into positive, negative, or neutral sentiments.

Documentation: Documenting the entire process, including the data collection, preprocessing, exploratory data analysis, model selection, feature extraction, model training, model evaluation, and model deployment.

Deliverables:

The following deliverables will be provided as a part of this project:

1. Codebase: The codebase will include Python scripts for data preprocessing, exploratory data analysis, model selection, feature extraction, model training, and model evaluation.

2. Trained Model: The trained model will be provided to classify new airline reviews into positive, negative, or neutral sentiments.

3. Documentation: The documentation will include a detailed description of the entire process, including the data collection, preprocessing, exploratory data analysis, model selection, feature extraction, model training, model evaluation, and model deployment.

4. Presentation: A presentation will be provided to explain the project, including the problem statement, dataset, preprocessing steps, model selection,

## Project Modules

The project can be divided into the following modules:

Data Collection:

This module involves collecting the Airline Twitter Sentiment dataset from Kaggle or any other reliable source.

Data Preprocessing:

This module involves preprocessing the text data by removing mentions, URLs, special characters, punctuations, stopwords, and stemming. It also involves converting the text data into a format that can be used by the Machine Learning algorithms.

Exploratory Data Analysis:

This module involves exploring the dataset to gain insights into the data, such as the distribution of sentiments, the frequency of words, and the most common reasons for negative sentiments. This analysis helps in understanding the dataset and identifying any issues or biases that may affect the Machine Learning model.

Model Selection:

This module involves selecting the appropriate Machine Learning algorithm for Sentiment Analysis. Some of the popular algorithms used for Sentiment Analysis are Logistic Regression, Naive Bayes, Random Forest etc. The selection of the algorithm depends on the size of the dataset, the complexity of the problem, and the performance of the algorithm on similar tasks.

Feature Extraction:

This module involves extracting features from the preprocessed text data using CountVectorizer or TF-IDFVectorizer. CountVectorizer converts a collection of text documents into a matrix of token counts, while TF-IDFVectorizer assigns weights to the terms based on their importance in the document.

Model Training:

This module involves splitting the dataset into training and testing sets and training the model on the training set. The model is trained on the features extracted from the preprocessed text data using the selected Machine Learning algorithm.

Model Evaluation:

This module involves evaluating the performance of the model on the testing set using metrics like accuracy, precision, recall, and F1-score. These metrics help in understanding the performance of the model and identifying any areas that need improvement.

Model Deployment:

This module involves deploying the trained model to classify new airline reviews into positive, negative, or neutral sentiments. The model can be deployed as a web application or an API, depending on the requirements of the client.

Documentation:

This module involves documenting the entire process, including the data collection, preprocessing, exploratory data analysis, model selection, feature extraction, model training, model evaluation, and model deployment. The documentation helps in understanding the process and replicating the project in the future.

The above modules can be implemented using Python programming language and various libraries such as Pandas, NumPy, Scikit-learn, and NLTK.

# 3. ER- Diagram

An ER-Diagram (Entity-Relationship Diagram) is not directly applicable to sentiment analysis projects, as it is typically used for visualizing database schemas and their relationships. However, we can assume that there may be a database component involved in this project, where the data collected from airline reviews is stored and accessed for analysis.

In that case, a possible ER-Diagram for this project could include the following entities:

1. Airline Reviews: This entity represents the main source of data for sentiment analysis. It includes attributes such as review text, sentiment label (positive, negative, or neutral), airline name, reviewer name, review date, etc.
2. Airlines: This entity represents the airlines being reviewed. It includes attributes such as airline name, airline code, country of origin, etc.
3. Reviewers: This entity represents the reviewers who provide the reviews. It includes attributes such as reviewer name, reviewer location, reviewer profile picture, etc.
4. Sentiment Labels: This entity represents the sentiment labels assigned to each review. It includes attributes such as sentiment label ID, sentiment label name (positive, negative, or neutral), etc.
5. Analysis Results: This entity represents the results of sentiment analysis performed on the reviews. It includes attributes such as review ID, sentiment label ID, prediction score, prediction probability, etc.

The relationships between these entities could be as follows:

* Airline Reviews and Airlines: Many-to-One relationship (Many airline reviews can be associated with one airline)
* Airline Reviews and Reviewers: Many-to-One relationship (Many airline reviews can be associated with one reviewer)
* Airline Reviews and Sentiment Labels: Many-to-One relationship (Many airline reviews can be associated with one sentiment label)
* Airline Reviews and Analysis Results: One-to-One relationship (One airline review can have one sentiment analysis result)

Overall, the ER-Diagram helps in understanding the structure of the database that stores the airline reviews and their related information, and how the different entities are related to each other

**4.Technologies to be used**

## Software Platform

1. **Front-end**

Jupyter notebook

1. **Back-end**

Python, Machine learning

## Hardware Platform

8GB RAM, 256gbSSD, Chrome etc.

# 5. Advantages of this Project

There are several advantages of using sentiment analysis with machine learning on airline reviews:

Improved Customer Satisfaction: By analyzing customer reviews, airlines can identify areas where they need to improve their services, and take corrective action to improve customer satisfaction.

Competitive Advantage: By analyzing customer sentiment, airlines can gain insights into their competitors' strengths and weaknesses, and use this information to gain a competitive advantage.

Cost-effective: Sentiment analysis with machine learning can help airlines automate the process of analyzing customer reviews, reducing the need for manual analysis and making the process more cost-effective.

Real-time Insights: Sentiment analysis with machine learning can provide real-time insights into customer sentiment, allowing airlines to respond quickly to customer feedback.

Data-driven Decision Making: By analyzing customer reviews using sentiment analysis, airlines can make data-driven decisions based on customer feedback, rather than relying on intuition or guesswork.

Improved Brand Reputation: By analyzing customer sentiment, airlines can identify negative reviews and respond to them in a timely and effective manner, improving their brand reputation.

Overall, sentiment analysis with machine learning can provide valuable insights into customer sentiment, allowing airlines to improve their services, gain a competitive advantage, and make data-driven decisions based on customer feedback.

# 

6. Future Scope and further enhancement of the Project

The sentiment analysis project on airline reviews using machine learning has several future scope and enhancement possibilities, including:

1.Multi-lingual Support: Currently, the sentiment analysis project only analyzes reviews in English. However, the project can be enhanced to support multiple languages, which will help airlines understand customer sentiment from different parts of the world.

2.Sentiment Classification Improvement: The current sentiment analysis project is based on a binary classification approach that categorizes reviews into positive or negative sentiments. However, the project can be enhanced to include a more nuanced sentiment classification system, which can categorize reviews into a wider range of sentiments.

3.Sentiment Analysis on Multimedia Data: Currently, the sentiment analysis project only analyzes text reviews. However, the project can be enhanced to analyze multimedia data such as images, videos, and audio, which will provide a more comprehensive understanding of customer sentiment.

4.Real-time Monitoring: The project can be enhanced to provide real-time monitoring of customer sentiment, allowing airlines to respond quickly to customer feedback.

5.Integration with Social Media Platforms: The sentiment analysis project can be enhanced to integrate with social media platforms, such as Twitter, Facebook, and Instagram, to analyze customer sentiment on these platforms.

6.Sentiment Analysis on Specific Services: The project can be enhanced to analyze customer sentiment on specific services provided by the airlines, such as in-flight services, check-in process, and baggage handling.

Overall, there are several opportunities for further enhancement and development of the sentiment analysis project on airline reviews using machine learning, which can help airlines improve customer satisfaction, gain a competitive advantage, and make data-driven decisions based on customer feedback.

**7.Team Details**

| **Project Name & ID** | **Course Name** | **Student ID** | **Student Name** | **Role** | **Signature** |
| --- | --- | --- | --- | --- | --- |
| SENTIMENT ANALYSIS USING MACHINE LEARNING | MAJOR-PROJECT | TCA2056023 | TANYA PAL | PROGRAMMER |  |
| TCA2056014 | MANISHA SINGH | DOCUMENTATION |  |

# 

8. Conclusion

In conclusion, sentiment analysis using machine learning on airline reviews is a powerful tool that can help airlines gain valuable insights into customer sentiment. By analyzing customer reviews, airlines can identify areas where they need to improve their services, gain a competitive advantage, and make data-driven decisions based on customer feedback.

The sentiment analysis project on airline reviews using machine learning has several advantages, including improved customer satisfaction, cost-effectiveness, real-time insights, data-driven decision making, and improved brand reputation. Furthermore, there are several future scope and enhancement possibilities, including multi-lingual support, sentiment classification improvement, sentiment analysis on multimedia data, real-time monitoring, integration with social media platforms, and sentiment analysis on specific services.

Overall, sentiment analysis using machine learning on airline reviews is a valuable tool for airlines to improve their services and customer satisfaction. As the technology advances, there will be more opportunities to enhance and develop the project, which will help airlines stay competitive and provide better services to their customers.

# 10.References

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1-2), 1-135.
2. Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011, June). Sentiment analysis of twitter data. In Proceedings of the workshop on languages in social media (pp. 30-38).
3. Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the conference on empirical methods in natural language processing (EMNLP) (Vol. 1631, pp. 1642-1654).
4. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-based systems, 89, 14-46.
5. Wang, X., Wei, F., Liu, X., Zhou, M., & Xu, K. (2016, August). Chinese microblog sentiment analysis based on HowNet and dependency parsing. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 2207-2212).
6. Hu, M., & Liu, B. (2004, July). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).
7. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1480-1489).
8. Lee, S. M., & Kim, D. H. (2016). A study on the classification of online customer reviews using a convolutional neural network. Journal of Intelligence and Information Systems, 22(4), 53-65.
9. Mishra, S., Kallakuri, P., & Adhikari, A. (2019). Deep Learning Approach for Sentiment Analysis of Airlines Tweets. In 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN) (pp. 183-187). IEEE.
10. Makhija, A., & Garje, G. R. (2021). Sentiment Analysis of Airline Reviews Using Machine Learning. In Proceedings of 3rd International Conference on Advances in Computational Intelligence and Informatics (pp. 345-351).

SCREENSHOT OF CODE















