Machine learning based Social Media Sentiment Analysis for Predicting Flight Delays in US Airlines

Khlood Alamoudi
Department of Information system and technology ,University of Jeddah
Saudi Arabia, Jeddah
khxxl009@gmail.com

Joud Aljehani
Department of Information system and technology ,University of Jeddah Saudi
Saudi Arabia, Jeddah
Aljehani.ajoud@gmail.com

Sadeem Alzahrani
Department of Information system and technology ,University of Jeddah
Saudi Arabia, Jeddah
Sadeem.s.az@hotmail.com

Muntaha Aldhari
Department of Information system and technology ,University of Jeddah Saudi
Saudi Arabia, Jeddah
Muntaha_12@hotmail.com

Rawan Hassoubah
Department of Information system and technology ,University of Jeddah
Saudi Arabia, Jeddah
rshassoubah@uj.edu.sa

***Abstract*— social media is crucial to understanding customers' feelings, emotions, and thoughts about their experiences anywhere. Leveraging these sentiments on platforms to develop and improve strategies for a company or organization will greatly help in making decisions that lead to the company's success and continuity. This paper proposes solutions to the problem of flight delays by presenting a comprehensive sentiment analysis and discovering the problems causing negativity within customers' tweets in twitter social media platform. Our study have utilizes a machine learning based models to predict flight delays within US airlines. The ML models along with (sentiment analysis) and (…..) used to analyze and predict the delays efficiently. It has shown an effective prediction with (value) (score). The proposed solution has been compared to current approaches and shown to be effective to improve customer satisfaction and to enhance the operational efficiency within US airlines and any other airline companies world wide.**

Keywords—social media analysis, sentiment analysis, machine learning, customer satisfaction.

I. INTRODUCTION

Social media platforms, especially Twitter, have become a significant source for understanding customers' feelings and satisfaction levels. Through these platforms, users express their emotions openly, making sentiment analysis a practical method for customer behavioral analysis and forecasting concerns. This process helps improve decision-making within an organization and fosters continuous improvements and developments.

In the airline industry, a crucial concern for customers is the airline's commitment to scheduled flight times. Flight delays cause significant dissatisfaction among customers and lead to concerns that result in losses for the airline industry. It is estimated that flight delays cost airlines approximately $28 billion annually due to operational inefficiencies, reimbursements, and lost customer loyalty. Many issues contributing to airline losses are directly linked to customer dissatisfaction. Therefore, understanding and managing customer satisfaction effectively can help airlines reduce losses and identify optimal solutions.

The act of identifying and classifying opinions in order to determine their sentiment polarity (whether positive, negative, or neutral) is called sentiment analysis, sometimes referred to as opinion mining. It helps businesses analyze consumer input and public opinion on social media. To manage the vast and varied amounts of data on sites like Twitter, sophisticated techniques like machine learning and deep learning are used [5].

The main objective of this study is to conduct sentiment analysis of US Airlines customers on the Twitter platform. The goal is to determine how sentiments correlate with the causes of flight delays. Using machine learning techniques, this study processes large volumes of Twitter data to uncover patterns in customer sentiment and its association with delays. Ultimately, the study seeks to build a machine-learning model that can predict flight delays, providing airlines with valuable insights to enhance customers satisfaction and operational efficiency.

This paper is structured as follows: the first section, presents the importance of leveraging social media sentiment analysis for the purpose of flight delay predictions concerning US airlines, identifying the objectives and relevance of the study. The second section, reviews related work, summarizing previous work on sentiment analysis related to the aviation industry in general. Next the third one, provides the problem statement, focusing on how hard it is to determine the delayed flights using traditional techniques. The methodology description including the dataset, preprocessing, and modeling are carried out in the fourth section. Later, An in-depth analysis of data for sentiment trends and their relation with flight delays forms the content. The sixth section describes the results obtained with the predictive model and their effectiveness compared to the current approaches. Lastly, A conclusion is presented along with the future work and recommendations through enhancement and new directions towards the refinement of the sentiment-based prediction model for airline operations.

II. RELATED WORK

To measure perceptions of airline companies, the authors of this study D. D. Das, et al in [1] examined sentiment on Twitter data, as the study included Emirates Airlines, United Airlines, and Jet Airways. The study focused

on classifying tweets into negative, positive, and neutral. The Naive Bayes algorithm was used to classify texts. Among the 1,298 tweets related to the United Airlines event, 395 (30.5%) were unfavorable, 187 (14.4%) were neutral, and 716 (55.1%) were positive. The researchers used tools such as R Studio in the investigation and Rapidminer to analyze and clean the data, which contributed to increasing classification accuracy by removing extraneous words.Based on the results of the Naive Bayes model, the tweets were classified with an accuracy of 72%. This was tested on other models, including support vector machines (SVM) and k-Nearest Neighbors (k-NN), and in comparison; SVM results showed a greater accuracy of 78%, while k-NN obtained 75%. These results indicate that the best model that gave the highest percentage of accuracy, the Support Vector Machine (SVM) model, is the most successful in classifying tweets. According to the study, Twitter sentiment analysis is an effective way to learn about customer opinions, but there is room to improve the results by using more data and experimenting with other classification models to obtain more comprehensive and accurate results.

This study by Y. Wan et al in [2] on sentiment analysis of airline services using Twitter data indicates that a majority voting-based sentiment classification system was used. The study used six individual classification algorithms: Naive Bayes, Support Vector Machine (SVM), Bayesian Network, C4.5 Decision Tree, and Random Forest. The researchers developed a new algorithm based on the clustering method that combines the results of these algorithms. They used a dataset of 12,864 tweets and added 10-fold model validation to verify the validity of the models. The results showed that the accuracy of the Naive Bayes algorithm was 70%, SVM 85%, Bayesian Network 76%, C4.5 Decision Tree 73%, Random Forest 82%, while the clustering algorithm method achieved an accuracy of 88%. The results indicate the success of the ensemble method over all individual algorithms, which enables airlines to improve their services based on customer feedback to contribute to service development and improve marketing strategies to increase effectiveness.

This research uses Twitter data to analyze sentiment related to US airlines using a voting classifier that combines logistic regression (LR) and stochastic gradient descent classifier (SGDC). LR achieved an accuracy of 0.789 using TF and 0.791 using TF-IDF. Several algorithms were used to support the study, including Decision Tree Classifier (DTC), which achieved a score of 0.754, Support Vector Classifier (SVC) with 0.765, Random Forest (RF) with 0.783, Gaussian Naive Bayes (GNB) with 0.732, AdaBoost (ADB) with 0.761, Gradient Boosting Machine (GBM) with 0.778, and Extra Tree Classifier (ETC) with 0.775. The results showed that the combined classifiers achieved higher numbers, indicating their superiority compared to the individual classifiers, which strengthens the evidence of using TF-IDF to improve classification accuracy. This study enhances the understanding of customer opinions better, and improves the airlines' customer experiences, which helps them in making decisions that contribute to developing services and attracting the largest number of new customers.

This research uses Twitter data to analyze sentiment related to US airlines using a voting classifier that combines logistic regression (LR) and stochastic gradient descent classifier (SGDC). LR achieved an accuracy of 0.789 using TF and 0.791 using TF-IDF. Several algorithms were used to support the study, including Decision Tree Classifier (DTC), which achieved a score of 0.754, Support Vector Classifier (SVC) with 0.765, Random Forest (RF) with 0.783, Gaussian Naive Bayes (GNB) with 0.732, AdaBoost (ADB) with 0.761, Gradient Boosting Machine (GBM) with 0.778, and Extra Tree Classifier (ETC) with 0.775. The results showed that the combined classifiers achieved higher numbers, indicating their superiority compared to the individual classifiers, which strengthens the evidence of using TF-IDF to improve classification accuracy. This study enhances the understanding of customer opinions better, and improves the airlines' customer experiences, which helps them in making decisions that contribute to developing services and attracting the largest number of new customers. F. Rustam, et al in [3].

In this study A. Samah et al. in [4]. develop Malaysian airlines (AirAsia, Malaysia Airlines, Malindo Air) using Twitter data. The researchers collected data throughout 2019 for sentiment analysis. Sentiment analysis. They used the Naïve Bayes algorithm to classify the data, which achieved an accuracy of 93% for English and 91% for Malay. The dataset included 800,000 positive and negative tweets in English, 344,733 negative tweets, and 312,985 positive tweets in Malay. Data processing techniques such as TF-IDF, Bag of Words, and Word2Vec were used. The researchers confirmed that the results of the proposed system enhance the provision of an effective interface for sentiment analysis, as the system usability reached 94.7%, which highlights the effectiveness of the system in understanding public opinion about Malaysian airlines.

III. PROBLEM STATEMENT

Flight delays are one of the major problems of airlines, causing disruptions in operations, increased costs, and reduced passenger satisfaction. In order to solve this problem, our research started with social media data, tweets in particular, to understand passenger sentiment. From the sentiment analysis, a sizeable portion of the negative comments by passengers was on flight delays. This underscored the need to exploit sentiment data in enhancing delay prediction capability.

Realizing the limitation of depending on sentiment data alone, we expanded our approach by developing a machine learning model using a secondary dataset that contained structured information about flights, such as historical flight schedules, weather conditions, and airline-specific factors. This approach integrates unstructured social media data with structured datasets, combining real-time public feedback with reliable operational metrics.

The advantages of this method are threefold: it gives airlines a more precise way to predict flight delays, makes resource allocation and scheduling more feasible, and allows the airline real-time insight into passenger concerns. It enhances

decision-making processes, reduces costs associated with delays, and ultimately improves overall passenger satisfaction.

IV. METHODOLOGY

This section presents the methodological steps that have been used during our work. Shown in fig.# the main steps that have been involved.

A. Dataset

Two datasets were used in this study: Twitter airline Sentiment Dataset and Flight Delay Prediction Dataset

- *Twitter airline Sentiment Dataset*

The study uses the **Twitter Airline Sentiment** dataset from Kaggle, featuring customer feedback from tweets since February 2015. Sentiments are classified as positive, neutral, or negative, with negative feedback including reasons like "late flight" or "rude service." This dataset helps analyze customer emotions and their links to flight delays, identifying patterns and associations.

- *Flight Delay Prediction Dataset*

The second dataset, ***flight_delay_prediction_data,*** contains 6,001 rows and 11 columns with features like **Delay Status, Weather Conditions,** and **flight metadata** (e.g., carrier, airport). It was used to train machine learning models for predicting flight delays, aiding airline operators in decision-making.

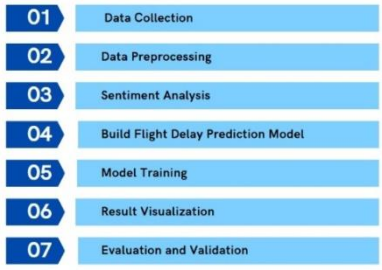

Figure 1 – steps of work

B. Data Processing

Cleaning, addressing missing values, encoding category variables, and normalizing numerical features were all part of the preprocessing of the data.

C. Sentiment Analysis

In this study, we analyzed Twitter data related to airlines, focusing on customer sentiment toward these companies. The analysis revealed several important insights into how flight delays affect passenger sentiment.

1) General Sentiment Distribution

The sentiment analysis showed that the majority of tweets were negative, with negative sentiments accounting for over 80% of the total dataset, as illustrated in **Figure 1**. This suggests that a large portion of passengers were dissatisfied with their experiences with airlines

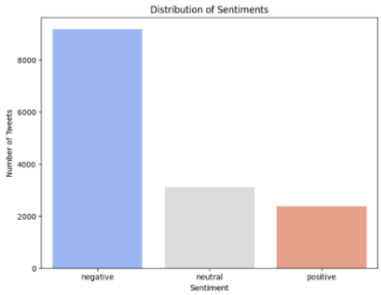

Figure 2: *Sentiment Distribution*

A sentiment trend analysis, based on daily tweets over a specific period, showed a dominance of negative sentiments. Particularly, negative sentiments peaked on certain days, such as **February 22**, indicating that passengers were more vocal about their dissatisfaction on these dates. This spike may have been triggered by negative experiences, including flight delays or cancellations.

2) Geospatial Distribution of Sentiments

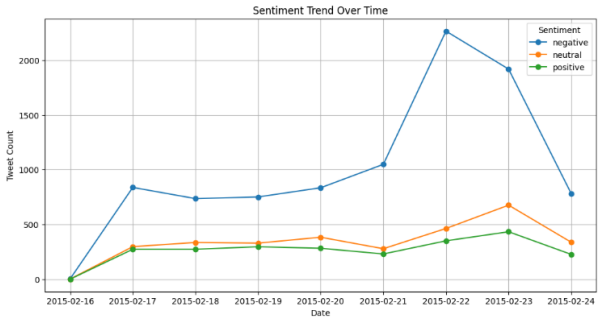

Figure 3: Sentiment Trend

An interactive map visualized the geographic distribution of tweets, highlighting regions with high concentrations of tweets, notably in cities with major airports such as New York, Chicago, and Los Angeles. This geospatial distribution provides insights into how sentiment varies by location and reflects the performance of local airports and airlines.

Figure 4: Geospatial Distribution of Sentiments

3) Retweet Analysis and Network Diagram

A network diagram was created to visualize the relationship between users and airlines through retweet analysis. The

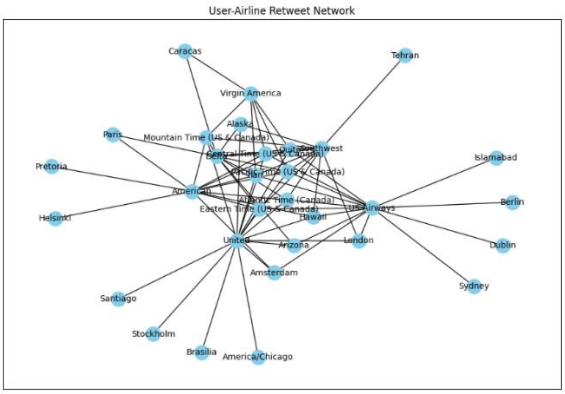

Figure 5: Network Diagram

diagram, where users and airlines are represented as nodes, reveals the social patterns and interactions surrounding customer feedback. The analysis allows us to observe how negative or positive sentiments impact an airline's reputation and how closely users are connected to the airlines in terms of sentiment expression.

4) Top Reasons for Negative Sentiment

The analysis identified **flight delays** as the leading cause of negative sentiment, with nearly **2,500 tweets** mentioning delays, as shown in **Figure 2**. **Customer service** was the second most frequently cited reason for negative sentiment. This confirms the significant impact that delays and poor customer service have on passengers' experiences with airlines.

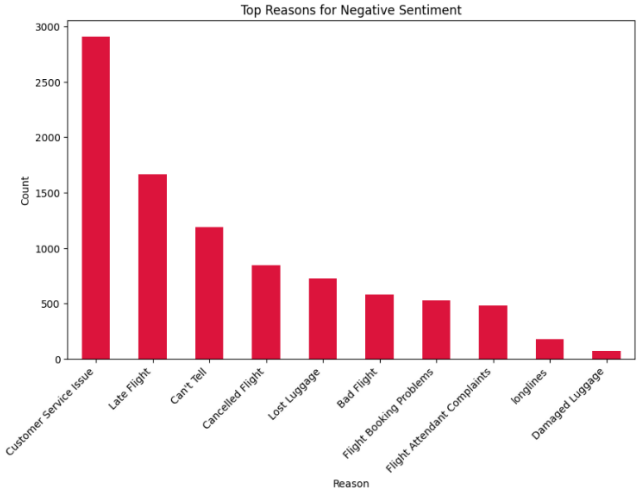

Figure 6: Reasons for Negative Sentiment

5)Word Cloud for Negative Sentiments

The word cloud for negative tweets highlights the most frequent terms associated with dissatisfaction, such as "flight", "delayed", "customer service", "cancelled", and various airline names. The prominence of terms like "flight" and "delayed" indicates that passengers are primarily concerned with delays and poor service. This word cloud effectively reinforces the recurring issues passengers face, centering on delays and service-related problems as key drivers of negative sentiment.

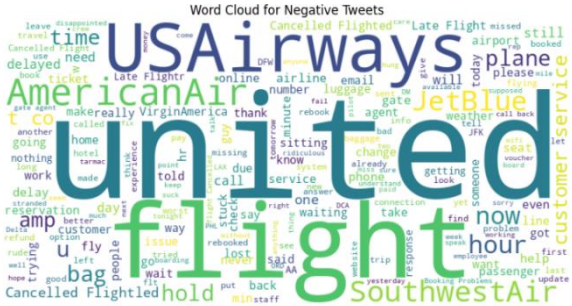

Figure 7: Word Cloud for Negative Sentiments

Given that flight delays are a major factor contributing to negative sentiment, we decided to develop a flight delay predictor based on sentimental analysis results. By leveraging machine learning models, we aim to predict flight delays and provide airlines with advance warnings. This proactive approach would allow airlines to take necessary actions to minimize the impact of delays on passengers, thereby improving customer satisfaction and operational efficiency.

D. . Building Flight Delay Prediction Model

A predictive model is designed to estimate flight delays based on input features such as weather conditions, airline schedules, and historical delays.

E. Model Training

Trained the following machine learning models:

1- Logistic Regression (LR)
2- Random Forest Classifier (RF)
3- XGBoost
4- Support Vector Machine (SVM)
5- K-Nearest Neighbors (KNN)

F. Visualization of Results

To assess the model's prediction accuracy, the model's performance was examined and represented using a confusion matrix. Additionally, the Sentiment Distribution is used in sentiment analysis to examine the distribution of emotions, including negative, positive, and neutral ones. A word cloud is also utilized to highlight the main themes and factors that lead to unfavorable attitudes. To illustrate the frequency of factors causing adverse reactions, a bar chart was made. The relationship between the several factors driving aircraft delays was depicted using a network diagram, and regional trends and patterns in delays and sentiments were extracted using a geospatial distribution mapping.

G. Model Evaluation and *Validation*

The performance of the model is evaluated using accuracy, precision, recall, and F1 score metrics, followed by validation for its reliability and robustness.

V. RESULTS AND DISCUSSION

A. Performance of Machine Learning Models

A predictive model was developed to analyze the likelihood of flight delays based on sentiment data extracted from Twitter tweets reflecting passengers' sentiments toward flights. To evaluate the performance of the models, three main machine learning algorithms—Logistic Regression (LR), Random Forest Classifier (RF), and XGBoost—were applied and compared using standard evaluation metrics: Accuracy, Precision, Recall, and F1-score. The implementation was performed through scikit-learn, with performance measured based on these metrics.

B. Experimental Setup

1)Preprocessing and Feature Engineering

- **Defining Target and Features**:
  The target variable, **y**, was defined as the **Delay Status**, indicating whether a flight was delayed or not.
  - **Categorical Columns**: These include **Departure Airport**, **Arrival Airport**, **Airline**, **Aircraft Type**, and **Weather Conditions**.
  - **Numerical Columns**: These include **Passenger Count** and **Distance (miles)**. These features were combined into a feature matrix.
- **Encoding Categorical Variables**: Categorical variables were transformed using **ColumnTransformer**, converting them into binary vectors suitable for machine learning algorithms.
- **Splitting the Dataset**: The feature matrix **X_encoded** and target variable **y** were split into a training set (80%) and a test set (20%) using the **train_test_split** function, with a **random_state** of 42 to ensure reproducibility.
- **Balancing the Training Dataset**: To address class imbalance, the **Synthetic Minority Oversampling Technique (SMOTE)** was applied, synthesizing additional examples of the minority class (flight delays).
- **Visualization**: A count plot was used to visualize the resampled target variable **y_resampled**, confirming the balance between delayed and on-time flights after applying SMOTE.

2) Implemented Models

- **Logistic Regression (LR)**: A linear classification model used for binary classification of flight delays.
- **Random Forest Classifier (RF)**: An ensemble model combining multiple decision trees to improve accuracy and handle feature interactions.
- **XGBoost**: A gradient boosting machine known for its effectiveness with tabular data and imbalanced datasets.
- **Support Vector Machine (SVM)**: A classification algorithm that finds the best hyperplane to distinguish between classes, using kernels to handle nonlinear relationships.
- **K-Nearest Neighbors (KNN)**: A simple algorithm that classifies data points based on the majority vote of their nearest neighbors, relying on distance metrics.

3) Evaluation Metrics

The models were evaluated using the following metrics:

- **Accuracy**: The percentage of correct predictions made by the model. Accuracy = TP + TN / Total where **TP** = True Positives, **TN** = True Negatives, and **Total** = Total predictions.
- **Precision**: The percentage of positive predictions that were correct. Precision = TP / (TP + FP) where **FP** = False Positives.
- **Recall**: The percentage of actual positives correctly identified by the model. Recall = TP / (TP + FN) where **FN** = False Negatives.
- **F1-Score**: The harmonic mean of precision and recall.
  (Precision * Recall) / (Precision + Recall) = 2 * F1 score. The F1-score helps balance the tradeoff between precision and recall, particularly in the case of imbalanced datasets.

C. Results

The test dataset, consisting of 20% of the data, was used to evaluate the models. The evaluation results are shown in the table below:

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	78.52%	81%	78%	78%
XGBoost	79.93%	83%	80%	79%
Logistic Regression	37%	37%	37%	37%
KNN	58%	58%	58%	57%
SVM	37%	37%	38%	37%

D. Observations

- **Best Models**: The **Random Forest** and **XGBoost** models performed best, with **XGBoost** slightly outperforming **Random Forest** in terms of accuracy and precision.
- **KNN**: Achieved moderate accuracy but struggled with recall for the **On Time** class. This may be due to the KNN algorithm's sensitivity to class imbalances and distance metrics.
- **Logistic Regression and SVM**: Both models showed poor overall performance, with an accuracy of 37%, which suggests they are not suitable for this dataset without additional optimization.

E. Discussion

1)Model Comparison

- **Random Forest**: This model achieved a good performance with an accuracy of **78.52%**. It showed relatively balanced precision and recall across both classes, benefiting from its ensemble nature. However, it was slightly outperformed by XGBoost.

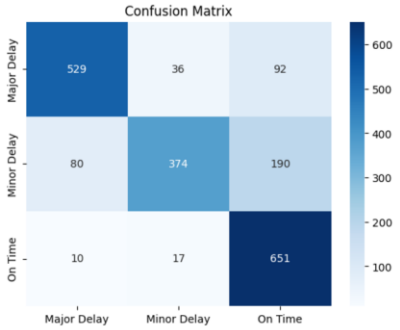

Figure 8: Confusion Matrix for Random Forest

- **XGBoost**: **XGBoost** achieved an accuracy of **79.93%**, slightly outperforming Random Forest. It demonstrated higher precision and recall in certain classes, showing its strength in handling complex data patterns and imbalanced datasets.

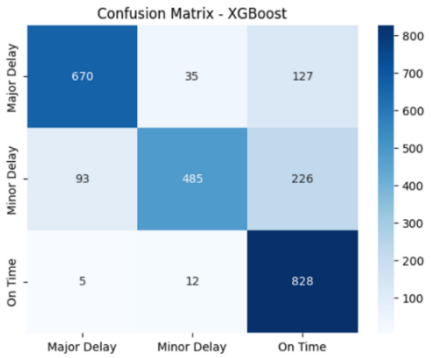

Figure 9: Confusion Matrix for XGBoost

- **Logistic Regression**: Logistic Regression struggled with this dataset, obtaining only a **37% accuracy**. Its limitations in handling non-linear relationships likely contributed to its poor performance.

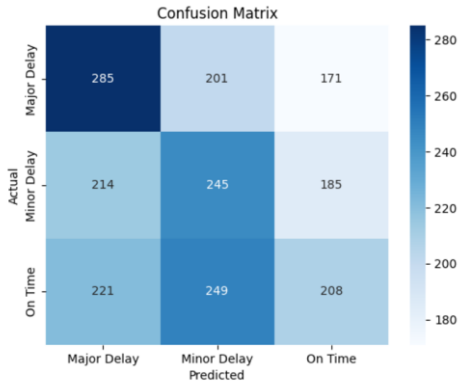

Figure 10: Confusion Matrix for Logistic Regression

- **KNN**: **KNN** performed moderately with an accuracy of **58%**, but it was sensitive to class imbalances, resulting in poor recall for the **On Time** class. KNN's reliance on distance metrics likely impacted its performance on this imbalanced dataset.

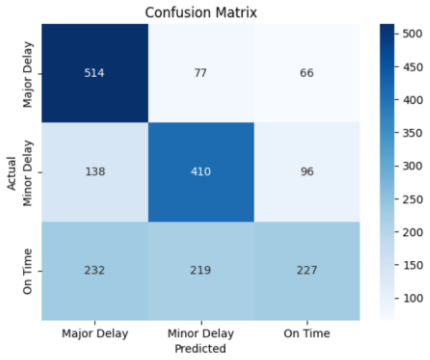

Figure 11: Confusion Matrix for KNN

- **SVM**: Like Logistic Regression, **SVM** performed poorly, with an accuracy of **37%**. The complexity of the dataset likely made it difficult for SVM to find an effective separating hyperplane.

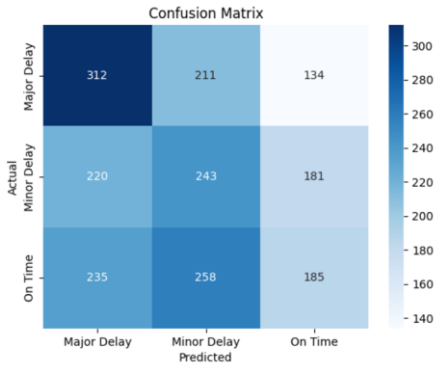

Figure 12 Confusion Matrix for SVM

2)Challenges

- **Class Imbalance**: One major challenge in this analysis was the class imbalance, with a disproportionate number of "On Time" flights. To address this, techniques like **SMOTE** were used to balance the classes in the training set. Despite this, some models (e.g., KNN and SVM) still struggled due to the inherent sensitivity to class distribution.

VI. FUTURE WORK

A. Real-Time Prediction

One area for improvement is enhancing the algorithms' applicability in **operational settings**, particularly for **real-time flight delay prediction**. Integrating the models with live data feeds (such as weather conditions, flight statuses, and air traffic) would improve their accuracy and responsiveness. By considering more **dynamic features**

(e.g., real-time flight information, delays from other flights, or immediate weather changes), the predictive models could be adapted to offer **timely alerts** and **proactive solutions** for flight delays. This would allow airlines to manage flight schedules more efficiently and passengers to receive accurate, real-time updates, leading to improved customer satisfaction.

B. Model Deployment

Another important step is **model deployment**, which involves integrating the developed machine learning models into **user-friendly applications**. These applications could be made available to both airlines and passengers. For airlines, a dedicated platform could use the models to provide **advance warnings** of potential delays, enabling proactive management of operations (e.g., adjusting staffing levels or rescheduling affected flights). For passengers, an app could notify them of **likely delays** and recommend alternatives, such as different flights or compensation options. The success of these models depends not only on their accuracy but also on their **ease of use** and **real-time capability**. Developing interfaces that are intuitive and easy to navigate for both passengers and airline staff will ensure broad adoption of the solution.

By addressing these future directions, the research could lead to more effective decision-making tools, enhancing both operational efficiency and customer experience in the airline industry.

VII. CONCLUSION

In conclusion, this paper highlights the significant role of social media, particularly Twitter, in shaping the decision-making processes of US Airlines, providing valuable insights into customer sentiment. The findings demonstrate that sentiment analysis of social media data is not only an effective method for understanding customer opinions but also plays a critical role in improving operational efficiency and enhancing customer satisfaction. The research proves that artificial intelligence, through sentiment analysis, can be a **turning point** in understanding customer behavior and making informed decisions.

This study successfully developed a model that explains the relationship between **customer sentiment** and **flight delays**, leveraging a **Twitter dataset** to predict delays based on the sentiment expressed in customer feedback. The analysis showed how sentiments in tweets could be categorized into **positive**, **neutral**, and **negative**, with negative sentiments strongly correlating with specific causes of delays, such as late flights.

Furthermore, the **Random Forest** and **XGBoost** models demonstrated the highest **accuracy** and **F1 scores**, proving their robustness in handling imbalanced datasets and effectively predicting flight delays. These results affirm the potential of machine learning models to enhance **operational decision-making** and improve overall **customer experience** in the airline industry.

As this research focused on a specific dataset, future studies could **expand the scope** by incorporating data from multiple platforms and further enhancing the model's generalizability. By doing so, airlines and other industries could develop more comprehensive solutions to address customer concerns and improve service delivery.

VIII. ACKNOWLEDGMENT

Special appreciation goes to **Ms. Rawan Hassubah** for her invaluable mentorship and unwavering support throughout this research. Her guidance and encouragement were crucial to the successful completion of this work. We also extend our sincere gratitude to the **Department of Information Systems & Technology, University of Jeddah**, for providing the necessary resources and a conducive environment for conducting this study. Additionally, we would like to thank all our colleagues for their contributions and collaboration, which greatly enriched the research process and outcomes.

IX. References

[1] D. D. Das, S. Sharma, S. Natani, N. Khare, and B. Singh, "Sentimental analysis for airline twitter data," in *IOP Conference Series: Materials Science and Engineering*, vol. 263, no. 4, p. 042067, Nov. 2017. IOP Publishing.

[2] Y. Wan and Q. Gao, "An ensemble sentiment classification system of twitter data for airline services analysis," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pp. 1318-1325, Nov. 2015. IEEE.

[3] F. Rustam, I. Ashraf, A. Mehmood, S. Ullah, and G. S. Choi, "Tweets classification on the base of sentiments for US airline companies," *Entropy*, vol. 21, no. 11, p. 1078, 2019.

[4] K. A. F. A. Samah, N. F. A. Misdan, M. N. H. H. Jono, and L. S. Riza, "The best Malaysian airline companies visualization through bilingual twitter sentiment analysis: a machine learning classification," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 1, pp. 130-137, 2022.

[5] Y. Liu et al., "Sentiment analysis on social media: A comprehensive study," IEEE Access, vol. 8, pp. 181154-181167, 2020.